International Journal of Science Education

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tsed20

Impact of Expert Teaching Quality on Novice Academic Performance in the Jigsaw Cooperative Learning Method

Roland Berger[a] & Martin Hänze[b]

[a] Physics Education Group, University of Osnabrück, Osnabrück, Germany

[b] Institute of Psychology, University of Kassel, Kassel, Germany

Published online: 08 Dec 2014.

International Journal of Science Education, 2015
Vol. 37, No. 2, 294–320, http://dx.doi.org/10.1080/09500693.2014.985757

Impact of Expert Teaching Quality on Novice Academic Performance in the Jigsaw Cooperative Learning Method

Roland Berger[a*] and Martin Hänze[b]
[a]*Physics Education Group, University of Osnabrück, Osnabrück, Germany;* [b]*Institute of Psychology, University of Kassel, Kassel, Germany*

We assessed the impact of expert students' instructional quality on the academic performance of novice students in 12th-grade physics classes organized in an expert model of cooperative learning ('jigsaw classroom'). The instructional quality of 129 expert students was measured by a newly developed rating system. As expected, when aggregating across all four subtopics taught, regression analysis revealed that academic performance of novice students increases with the quality of expert students' instruction. The difficulty of subtopics, however, moderates this effect: higher instructional quality of more difficult subtopics did not lead to better academic performance of novice students. We interpret this finding in the light of Cognitive Load Theory. Demanding tasks cause high intrinsic cognitive load and hindered the novice students' learning.

Keywords: *Cooperative learning; Instructional quality; Task difficulty; Cognitive load theory; Physics education; Quantitative research*

Introduction

Teaching by students is an important component of all cooperative learning models. In student 'expert' models, teaching can be structured in several different ways. Widespread 'expert' models include tutoring (Fantuzzo, King, & Heller, 1992), reciprocal teaching (Palincsar & Brown, 1984; Rosenshine & Meister, 1994), and several task-specialisation methods (Slavin, 1983, 1996) such as group investigation and the jigsaw learning technique.

The jigsaw method is a widely used cooperative learning method first proposed by Aronson in the 1970s (Aronson, 2002; Aronson, Blaney, Stephan, Sikes, &

*Corresponding author. Physics Education Group, University of Osnabrück, Barbarastr. 7, Osnabrück 49076, Germany. Email: r.berger@uos.de

Snapp, 1978). This form of group work involves students switching between different groups and acting as both expert teachers and novice students. Students first form 'expert groups' that are assigned a specific subtopic. Together, students within an expert group research and discuss the subtopic, and address questions and problems. Subsequently, these expert groups break up and the students recombine with 'experts' in other subtopics (from different expert groups) to form teaching groups. Each student in the group then teaches the rest of the group (novice students) his or her expert subtopic. In this way, the whole topic is taught.

Due to task specialisation, each member of a group is accountable for one unique part of the learning material. Novice students in the teaching groups are strongly dependent on the knowledge of experts (resource interdependence). Thus, the jigsaw classroom is based on positive interdependence, an element considered essential for student success in a cooperative learning environment (Cohen, 1994; Johnson & Johnson, 2009).

Nevertheless, several studies have demonstrated that novice students in the teaching groups are outperformed by expert students in subsequent tests (Hänze & Berger, 2007; Souvignier & Kronenberger, 2007). This performance difference between expert students and novice students might in part be traced back to the teaching expectancy that motivates expert students to study the learning material more intensively (Renkl, 1995), and consequently more time-consuming if time-on-task is not limited. Slavin, Hurley, and Chamberlain (2003) attributed the performance difference between expert students and novice students to task specialisation; students have limited exposure to material and so are highly dependent on the teaching quality of the expert, particularly if there are no supplementary resources (e.g. written explanations) to compensate for poor expert teaching quality (e.g. low-level verbal explanations) (Wittwer & Renkl, 2008). Thus, teaching quality is a critical factor in the success of expert models of cooperative learning.

In the present study, we focused on the relationship between the expert teaching quality and the academic performance of novices in the teaching groups. To the best of our knowledge, there are no studies investigating the interrelation between teaching quality assessed by external raters and novice academic performance in a science classroom organised according to the jigsaw cooperative learning model. A significant relation between these variables, while plausible, is not self-evident. For example, even high teaching quality is fruitless if novices are not able to process the information.

Buchs, Butera, and Mugny (2004) rated the behaviour of both expert and novice university students working in groups on a social psychology topic using a low inference scale that measured the time devoted to explanations and the number of responses to questions. The authors found that the number of ideas experts transmitted to novice students was linked to novice academic performance. More recently, Moreno (2009) audio-recorded small group discussions and classified statements according to three cognitive levels: retention, elaboration, and metacognition. In jigsaw teaching groups, a relatively small proportion of these statements were the most effective 'elaboration statements', but correlations between teaching quality and novice academic performance were not reported.

In this study we demonstrate that the relationship between teaching quality and academic performance of novice students strongly depends on the difficulty of the subtopic. We argue that this is due to the varying cognitive demands imposed by different subtopics and present empirical evidence to support this view. This finding strongly suggests that content selection is critical for the quality of learning in jigsaw classrooms and related expert models of cooperative learning.

Background

Teaching Quality in the Jigsaw Teaching Groups

In cooperative learning settings, the learning process crucially depends on the quality of the interaction between students. Slavin (1997) proposed the QuAIT model, which combines features of teaching quality considered indispensable for successful teaching. The model involves several factors that can be affected by teachers and encompasses all potential forms of classroom organisation, including cooperative learning. The elements of the QuAIT model are structural *Qu*ality, *A*ppropriateness, *I*ncentive, and *T*ime.

Structural *Qu*ality depends on how information is presented so that students can easily learn it. Information must be presented in an organised, orderly way. The teacher must frequently restate essential principles, remind students of previously learned material at relevant points in the lesson, and use frequent formal or informal assessments with immediate feedback to students. *A*ppropriateness is the degree to which the teacher prepares the students to learn new material by ensuring that they have the prerequisite skills and knowledge. Based on the classification proposed by Alexander, Kulikowich, and Schulze (1994), Renkl, Stark, Gruber, and Mandl (1998) highlighted that prior topic knowledge (i.e. the physics underlying the electron microscope) should predict post-test performance better than broader knowledge of physics (domain knowledge). Hence, assessment of the prior topic knowledge, that is, the specific knowledge that is relevant for understanding the topic, is recommended in order to construct a powerful predictor for post-test performance. *I*ncentive is the degree to which students are motivated to learn. According to the model of Wigfield and Eccles (2000), several subjective task values are assumed to directly influence achievement-related choices and performance. The 'intrinsic value' component of subjective task values is similar to the intrinsic motivation construct as defined by Deci and Ryan (2000). Intrinsic motivation is associated with the application of more effective deep level processing strategies (e.g. Schiefele, 1991), so this variable should be a significant predictor of post-test performance. *T*ime refers to the time allocated for a given lesson (matched to difficulty). Optimisation (not maximisation) of lesson speed (pacing) should allow a student to study even complex material without time pressure.

Based on the empirical results of Brophy and Good (1986), Weinert, Schrader, and Helmke (1989) outlined important facets of quality of instruction that are accepted by most researchers (Neumann, Kauertz, & Fischer, 2012). Among these are pace of instruction, the structuring of information to be learned, clarity of presentation, and proper feedback from the teacher. More recently, Seidel et al. (2007) provided empirical evidence that goal clarity and coherence impact students' perceptions of supportive learning conditions. In a large meta-analysis, Hattie (2003) substantiated the relevance of coherent teaching and useful feedback for successful learning.

Due to resource interdependence, the interactions between expert and novice students in the jigsaw teaching groups are characterised more by transmission of knowledge from expert to novice and less by a co-construction of knowledge within a truly cooperative working group (Moreno, 2009). Hence, it is reasonable to assume that not only students in a traditional direct instructional setting but also novices in the jigsaw teaching group benefit from structured and coherent teaching and an appropriate teaching pace.

Why is structured and coherent teaching crucial for successful learning? From the perspective of cognitive psychology, a logical order of presentation, emphasis on essential principles (through summaries or comprehension questions), and frequent feedback foster the development of a well-organised knowledge base. A coherent presentation focuses students' attention on the most important concepts, provides a framework to integrate new information into the knowledge base, and ultimately facilitates deeper understanding (Rosenshine & Stevens, 1986). In the conceptual language of cognitive load theory (CLT), appropriately structured lessons can free working memory capacity to promote information processing and integration (Sweller, 1988, 2010) if the pacing matches the learners' skills and task-related prior knowledge.

Based on these research results, in the present study we define the teaching quality construct by five characteristics. We assume an appropriate *teaching pace* if the expert student adapts the pace to the difficulty of the subject matter and gives appropriate time for discussion of difficult aspects. The expert students' *teaching is clear* if he or she clearly explains the goals of the unit, provides summaries of important facts, and distinguishes seminal from less important information. The expert students' *teaching is coherent* when he or she presents the material in an organised fashion, and explains how certain aspects relate to each other and to prior topic knowledge.

With regard to a proper feedback, we assessed characteristics that encompass metacognitive strategies. The expert student shows consideration for his group if he or she frequently asks if the novice students understand the material, and poses comprehension questions (*monitoring*). The group deals with comprehension problems appropriately if the expert and/or the teaching group addresses problems and tries to solve them (*regulation*).

Verbal explanations are a frequently used strategy in classroom teaching, tutoring, and peer-to-peer learning. The majority of research, however, shows that verbal explanations do not necessarily foster deeper understanding. A meta-analysis by Webb (1991) found that verbal explanations had a low impact on student performance. A crucial factor for the effectiveness of verbal explanations is the degree to which the explanation triggers elaboration (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Webb & Farivar, 1999). According to Webb, explanations are effective only if several requirements are fulfilled. Obviously, the explanation must be understood by the learner. The level of teaching must be appropriate, neither too difficult nor too easy (Slavin, 1997; Wittwer & Renkl, 2008), as superficial explanations will not stimulate high-level cognition (elaboration). Thus, while high teaching quality is an obvious necessity, it does not guarantee successful learning.

The Cognitive Demands of Learning Tasks

Low test performance may not necessarily result from poor-quality teaching (Ing, 2008). In light of CLT, material requiring high cognitive demand leads to high cognitive load and reduced learning (Paas, Renkl, & Sweller, 2003). If the learning material is highly demanding, even relatively high-quality teaching may not result in satisfactory performance by novices. This might partly explain the lower academic performance of novices compared to expert students in the jigsaw classroom.

Since the concept of 'cognitive demand' was first developed by Bloom (1956), a variety of taxonomies have been proposed to describe cognitive demand. Howe and Durr (1982) rated the difficulty of test items for the learning unit 'concept of the mole' (e.g. calculations involving Avogadro's number). They proposed a three-level category system based on Piagetian stages of development: late concrete operational, early formal operational, and late formal operational. Late concrete operational processing is required to perform calculations involving simple arithmetic operations, early formal operational processing requires inferences from actual experience, and late formal operational processing involves making inferences from models or theories. Edwards and Dall'Alba (1981) attempted to place the cognitive demand construct in a broader theoretical framework derived from a range of theories of learning and cognition. Cognitive demand was defined as the demand placed on cognitive capacity by several facets of the material or concept: complexity, openness, implicitness, and level of abstraction. The authors proposed a scale for measuring cognitive demand for test items in physics. In instructional science laboratories, Dreyfus (1986) identified factors that could determine the difficulty of the various phases in a typical research sequence (generation of hypotheses, performance of the experiment, analysis, and interpretation of data). The author found that the number of conceptual elements[1] plays a significant role in determining the cognitive demand.

These category systems were developed to assess specific, clearly defined contexts (e.g. tests or laboratory projects). In our study, we are faced with jigsaw tasks that are quite diverse. They involve activities such as running a simulation, comprehending the functioning of a detector, drawing the magnetic field lines on a figure, and conducting an experiment. These tasks do not proceed along a standard sequence of stages. Due to this complexity, we refer to the CLT to help define cognitive demand (Paas et al., 2003; Sweller, 1988; Sweller, Ayres, & Kalyuga, 2011). The CLT provides a sufficiently flexible framework to account for the deep level structure of tasks. In the following section, we discuss some basic aspects of CLT most relevant to the present study.

According to the CLT, the working memory load imposed by a task depends on the number of conceptual elements that must be processed simultaneously in working memory. The intrinsic cognitive load of learning material is defined by the number of conceptual elements and the interrelations between conceptual elements (element interactivity) that need to be learned. Material with low element interactivity can be understood by comprehension of one element at a time without consideration of other elements. Thus, low element interactivity allows for sequential learning, element by element. In contrast, to understand material with high element interactivity, the learner must process all of the elements of the task and their interactions simultaneously. Thus, material that consists of many interacting elements is harder to understand because the elements cannot be easily held simultaneously in working memory. Due to limited working memory, high intrinsic cognitive load engages memory resources that could otherwise be used to actively process the information (germane load). Thus, intrinsic cognitive load correlates with the number of interacting conceptual elements needed to solve a task (Brünken, Seufert, & Paas, 2010) and is a predictor of task difficulty (Schnotz & Kürschner, 2007). In contrast, a task that is characterised by many independent elements (e.g. the symbols for chemical elements) may be difficult because many symbols have to be learned, but it does not impose a heavy working memory load because there is no need for 'understanding' or relating conceptual elements (Sweller, 2010; Sweller, van Merrienboer, & Paas, 1998). For this reason, we will adopt the term 'cognitive demand' in order to characterise the cognitive load imposed by a task.

An elaborated knowledge base is characterised by a large number of schemas that incorporate multiple elements of information into a single element. Once a schema has been constructed, interacting elements are incorporated within it and do not need to be held individually within working memory. For this reason, intrinsic cognitive load is determined by an interaction between the nature of the learning material and the expertise (prior knowledge) of the learner. A well-developed task-related knowledge base frees working memory capacity and increases germane load, thus fostering the learning process. Hence, estimation of element interactivity by analysing the task requires assumptions concerning the degree of prior expertise (Chandler & Sweller, 1996).

Alternatively, if the learner has to take different sources of information into account in order to perform a task successfully, this may lead to increased cognitive load through the 'split-attention effect' (Sweller et al., 1998). When the task involves running an experiment or a computer simulation for example, the student has to read the manual, store this information in working memory, and subsequently search for appropriate referents in the experiment or the simulation. Thus, the student must split his or her attention between different sources and mentally integrate information from the manual and the experiment or simulation. This ongoing process of mental integration is likely to impose a heavy extraneous cognitive load. Enhanced extraneous cognitive load might hinder learning when dealing with material that has a higher level of intrinsic element interactivity (Sweller & Chandler, 1994).

Research Goals

The primary goal of the present study (Research goal 1) was to examine which variables predict novice academic performance as assessed by scores on post-test items evaluating comprehension of the subtopic taught by the expert. According to Slavin's QuAIT model, we hypothesise that novice academic performance correlates with teaching quality of experts, novices' intrinsic motivation, and their prior topic knowledge.

Second, we examined if the impact of instructional quality on novice academic performance depends on the cognitive demand of each subtopic (Research goal 2).

Methods

The Learning Unit

Our study is based on a learning unit in 12th grade physics selected by the following criteria. For the jigsaw classroom model, the topic had to be divisible into independent segments or subtopics. Furthermore, the topic had to be meaningful—that is, important for the 12th grade level. Hence, it was suggested that the students should learn about the principles of the scanning electron microscope (SEM). This topic is particularly suitable for teaching physics at the 12th grade level because the underlying physical principles (the motion of charged particles in electric and magnetic fields) encompass a major portion of the curriculum.

The following core concepts essential to grasp the physics of the SEM were taught by the teachers prior to the research study:

(1) Electric fields (capacitors, electron gun based on thermionic emission[2]).
(2) The magnetic field of a coil (particularly the 'left-hand rule').
(3) Motion of charged particles in a magnetic field (particularly the 'three-finger rule').

Thus, these core concepts were repeated in the context of the SEM, so the learning unit required only four hours.

The post-test evaluated four subtopics discussed in the jigsaw classroom (2 hours). These subtopics are described in Appendix 1. In addition, Appendix 1 discusses the relevance of prior knowledge to each jigsaw subtopic and describes the conceptual elements of each subtopic.

The relevance of prior topic knowledge from prior instruction for each subtopic is rated in Table 1. The relevance of prior topic knowledge is low for the subtopics *penetration depth* and *electron detector* and high for the subtopics *electron gun* and *beam deflection* (cf. second column in Table 1). Hence, the intensive study by the expert groups should lead to a larger performance gap with novices on the subtopics *penetration depth* and *electron detector* than the subtopics *electron gun* and *beam deflection*.

Table 1. Features of the subtopics

Subtopic	Relevance of prior topic knowledge	Visualisation media
Electron gun	High	Hitachi electron emission wire cathode
Beam deflection	High	Experiment (cf. Figure A1)
Penetration depth	Low	Computer simulation
Electron detector	Low	Schematic drawing of the detector

Resources for Teaching

Students in the expert groups learned their assigned subtopic using prepared texts (e.g. Appendix 2 worksheet 'How is the electron beam deflected?'). They were asked to study the material cooperatively in order to teach their peers in the teaching group. The experts were requested to formulate keywords as written help for later teaching. Adequate keywords should promote clear and coherent teaching. To address deficiencies in task-related prior knowledge, students had access to written help on the teachers' desk (e.g. Appendix 2 'Remember the electron beam deflection').

The expert students were encouraged to use visual media in the teaching groups (cf. Table 1). To visualise the *electron gun*, the students were provided with a Hitachi electron emission wire cathode. In order to visualise the *beam deflection*, students had to conduct an experiment where an electron beam in a vacuum tube has to be deflected by manipulating the electric current in a coil. In order to consolidate their knowledge, novice students were asked by the expert to apply the 'left-hand rule' and the 'three-finger rule' in a paper and pencil task. To perform this task, students had to transfer their knowledge from an experiment to a schematic drawing of an electron beam that passes the coil. The study on how *penetration depth* depends on both primary electron energy and atomic number is assisted by a computer simulation based on a Monte Carlo algorithm.[3] The principles of an *electron detector* were discussed using a schematic drawing of the detector. Adequate inclusion of this material should foster the clarity of experts' explanations in the teaching groups.

Participants, Design and Procedure

Nine grade 12 physics classes from 7 schools (a total of 129 students,[4] 87 males, 38 females, 4 students did not provide gender information) participated in the study. All 129 students acted as both an expert and a novice. The present report is part of a larger study conducted to compare two methods of cooperative learning, the jigsaw technique and the cyclical rotation method (Berger & Hänze, 2009). It includes additional unpublished analyses. The design of the learning unit is shown schematically in Figure 1.

In the physics hour before starting the lessons, students were tested for prior topic knowledge. The learning unit consisted of four school hours. First, basic information about the motion of electrons was introduced in two hours of direct instruction by the

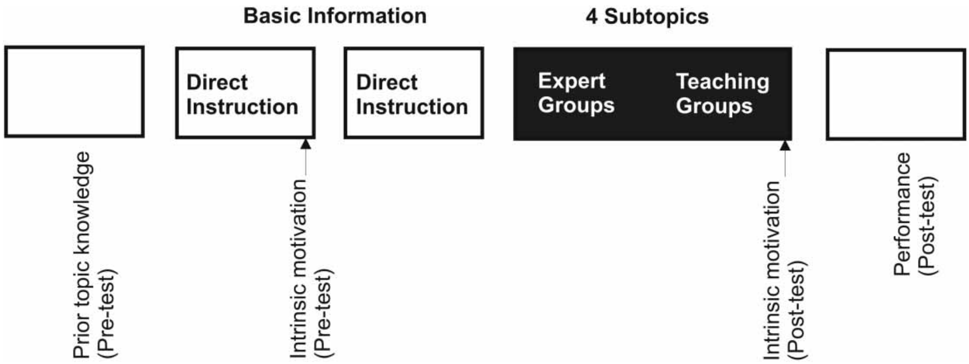


Figure 1. Study design. Each period (block) was one hour

teacher. At the end of the first hour, a questionnaire concerning intrinsic motivation was administered as a pre-test measurement. In the next two physics hours (a double period), students worked in the jigsaw classroom. Temporary expert groups were formed consisting of 3–5 students. Each group was assigned one of the subtopics (*electron gun*, *electron beam deflection*, *penetration depth*, or *electron detector*). Students were assigned randomly to specific expert and teaching groups by drawing cards. After the expert stage, the students joined teaching groups and taught each other their respective subtopics. No time limit was given to the students in either group. The expert group phase lasted about 30 minutes. The teaching group session lasted about 40 minutes. In order not to overstrain students, the intrinsic motivation questionnaire was not administered after every subtopic but rather after students were taught all four subtopics by the experts. The post-test of academic performance was given in the next physics hour several days after the jigsaw classroom. The conversations in the teaching groups were audio-recorded. Additionally, the experiment in the subtopic *electron beam deflection* was videotaped in order to facilitate the rating of teaching quality.

Instruments and Measures

According to our research questions, we assessed academic performance, teaching quality of experts, novices' intrinsic motivation, and their prior topic knowledge. Furthermore, we estimated cognitive demand of each subtopic according to rates by physics education specialist. Due to time constraints, no further control variables were assessed.

Intrinsic motivation. To measure intrinsic motivation, we used the German version of the self-report scale described by the Berger & Hänze (2009). Cronbach's alpha for internal consistency reliability was 0.69 for intrinsic motivation as assessed by three items (e.g. 'I was eager to learn the material.). Students responded to the items on a 5-point scale in which only the first and fifth points were anchored ('strongly disagree' to 'strongly agree').

Academic performance tests. Students took pre- and post-test in physics. The pre-test consisted of four open items that assessed students' prior topic knowledge of concepts deemed essential for understanding the physics of the SEM in the jigsaw classroom. The first item addressed knowledge about charged particles (charge of electrons and atomic nuclei as well as their interaction). Two items were related to the 'left-hand rule' and the 'three-finger rule', and one item probed students' understanding of the electron gun. One example item is 'Explain by a drawing how electrons in an electron tube are released and accelerated.' The post-test evaluated learning of the four subtopics taught by experts in the teaching groups. The post-test items and the scoring scheme can be found in Appendix 3.

For individual students, both the pre-test and post-test results correlated with their last grade earned in physics, with r values of 0.42 (pre-test) and 0.45 (post-test). The internal consistency as measured by Cronbach's alpha coefficient was 0.45 for the pre-test and 0.60 for the post-test, which we considered acceptable given the limited number of test items and the broad range of tested knowledge. The field-experimental character of the study did not allow for more comprehensive performance testing. Nonetheless, the mean inter-item correlations on the pre-test (0.18) and on the post-test (0.30) were within the range recommended by Clark and Watson (1995) for reasonable homogeneity.

Both tests were corrected by a physics education student with a Bachelor's degree in physics. Each item was scored based on a scoring scheme. The student was trained by one of the researchers beforehand. The interrater reliability between the student (primary rater) and the researcher was tested on a sub-sample of 23 tests. The intra-class correlation was 0.96.

Teaching quality. In the following, we briefly describe the items and the corresponding criteria used to assess teaching quality. The first two items are adapted from Herweg, Seidel, and Dalehefte (2005).

(1) 'The teaching is clear': The expert …
- … clearly states the goals (e.g. 'And now we address the question of how the penetration depth depends on the atomic number of the specimen material.');
- … provides or requests summaries of important facts (e.g. 'Repeat the two main factors that determine penetration depth.');
- … distinguishes critical from less important information (e.g. novice: 'What do the blue and pink bars on the screen mean?'/expert: 'This is not important.').

(2) 'The teaching is coherent': The expert …
- … presents the material in an organised, orderly way (e.g. 'We already discussed the dependence on the atomic number and come now to the energy dependence of penetration depth.');
- … explains how certain aspects relate to each other (e.g. 'Note the difference between gold and carbon, which we discussed earlier.');
- … relates new content to prior topic knowledge (e.g. 'That is precisely what we discussed yesterday').

(3) 'The teaching pace is appropriate': The expert …

- … adapts the pace to the difficulty of the subject matter;
- … gives appropriate time for discussion of difficult aspects.

Furthermore, we adopted two items that encompass metacognitive strategies:

(4) 'The expert shows consideration for his group' (monitoring): The expert …
- … frequently asks if the students understand the material (e.g. 'Did you understand this?');
- … poses comprehension questions (e.g. 'Why is the penetration depth of an electron smaller than its path length?').

(5) 'The group deals with comprehension problems appropriately'[5] (regulation): This is the case if the expert and/or the teaching group addresses problems and tries to solve them (e.g. expert: 'Electron penetration depth is much higher in titanium.'/ novice: 'But why?'/expert: 'This is due to the lower atomic number.').

Teaching quality was rated on a high inference level based on these five items. Each item was rated on a 5-point scale in which only the first and fifth points were anchored (strongly disagree' to 'strongly agree') by two advanced physics education students. The analysis unit for applying the high inference rating procedure was a complete expert teaching of a subtopic. Each period of expert instruction lasted several minutes (mean length 7.2 minutes, standard deviation 2.9 minutes).

The raters were familiarised with the items, corresponding criteria and examples as depicted earlier, and the underlying teaching quality construct prior to the rating process. The raters were encouraged to take notes while listening to the expert students' presentations. In case of doubt, they were free to rehear the audio recordings. The raters filled out the rating scale after listening to every single expert presentation in the teaching group.

For training purposes, two audio recordings of exert presentations for each subtopic were transcribed and subsequently rated independently by the physics education students. Rating differences were discussed with participation of one of the authors until agreement was reached.

To estimate the interrater reliability, 25% (32 out of 129) of the expert presentations were double-rated. The intraclass correlations (two-way random effects model, with single measure and absolute agreement) for the five teaching quality items were 0.62, 0.51, 0.72, 0.60, and 0.70. Considering the high inference scale, these values were acceptable. According to Cicchetti (1994), four items reflect 'good' agreement and 1 item reflects 'fair' agreement. Considering the whole five-item scale, intraclass correlation was 0.74.

Factor analysis revealed one factor with eigenvalue exceeding unity that accounted for 58% of explained variance. Hence, for each student expert, we computed the variable 'teaching quality' as the mean of the five-item scores. Cronbach's alpha for the scale was 0.82. Due to the good internal consistency, we do not analyse single aspects of the construct 'teaching quality'. The correlation between the teaching quality and the post-test academic performance of experts in their subtopic was 0.41. This significant result is in accordance with Fuchs et al. (1996), who showed experimentally that academically successful tutors gave better explanations. The moderate correlation reflects the fact that high-quality teaching does not depend exclusively on well-developed expert knowledge. Furthermore, task-related contributions (e.g. timely questions) from other group members may also influence teaching quality. The correlation between expert teaching quality and own post-test performance may be interpreted as an aspect of construct validity of the teaching quality rating system. Nevertheless, this should be interpreted with due caution because experts may benefit from explaining and responding to novices' questions (Roscoe & Chi, 2007; Webb, 1991), leading in turn to higher expert post-test performance and overestimation of the true correlation.

Cognitive demand. According to Brünken et al. (2010), the number of interacting elements needed in working memory to solve a task is an indicator of cognitive load. In order to assess the intrinsic cognitive load of the subtopics, we conducted 'unit of meaning-based concept mapping' in collaboration with seven physics education specialists. Each of the specialists holds a Master's Degree in physics and has experience teaching physics in high school as well as teaching physics education at university. Each subtopic was divided into 'units of meaning' by each physics education specialist. Subsequently, each specialist developed a concept map for each of these units of meaning. Concept mapping is regarded as a reliable technique to represent the knowledge structure of subject matter (Novak, 1998). For a better understanding of this procedure, the concept maps proposed by one of the participating specialist are depicted in Figure 2. The specialist divided the subtopic 'electron gun' in two units of meaning, (1) release of electrons off the cathode and (2) acceleration of free electrons, each with four conceptual elements: (1) electron, hot wire, thermal motion, and heating voltage and (2) electron, anode, acceleration voltage, and hot wire. The relationships between conceptual elements are also shown.

For each of the four subtopics, a list of conceptual elements was proposed by one of the researchers to the specialists, who were then free to add or omit elements. According to CLT, the number of conceptual elements depends on prior learning and concomitant schema acquisition. To take this into account, the core concepts that had been taught by the teachers prior to the study (cf. section 'The Learning Unit') were communicated to the specialists. The number of units of meaning was defined by the specialists.

This method of estimating cognitive demand accounts for the number of conceptual elements within one unit of meaning as well as the relationships between elements. The degree of element interactivity determines the degree to which the information taxes working memory. In line with CLT, we assume that a large number of elements per unit of meaning lead to high cognitive load. For a quantitative analysis, we used the mean number of elements per unit of meaning as an indicator of the cognitive demand of the subtopic. In the given example, the mean number of conceptual elements for the subtopic 'electron gun' is 4.0. Intraclass correlation (two-way random effects model, with average measure and absolute agreement) was 0.90.

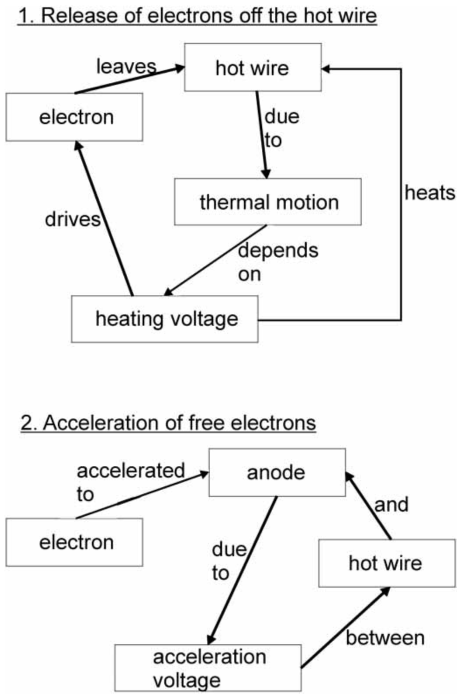


Figure 2. Sample concept maps for the subtopic 'electron gun' encompassing two units of meaning

Statistical Analyses

We first compared teaching quality between subtopics and academic performance between expert students and novice students. Second, we examined the correlations between novice academic performance and expert teaching quality, novice intrinsic motivation, and novice prior topic knowledge. These variables were then analysed simultaneously by hierarchical linear regression analyses with novice academic performance as the dependent variable. These regression analyses were run with HLM version 6 and all other analyses were conducted with SPSS version 18.

Results

Research Goal 1: Examination of Relations to Factors Influencing Novice Academic Performance

Teaching quality. Overall teaching quality, expressed as the mean of the 5 rating items for each of the subtopics, is depicted in Table 2 (standard deviations in brackets).

Analysis of variance demonstrated that the teaching quality depended on subtopic ($F(3, 98) = 6.54$; $p < .001$). Teaching quality was highest for the subtopic *electron detector* and lowest for the subtopic *penetration depth*.

Table 2. Mean teaching quality in teaching groups (standard deviations in brackets)

Subtopic	Teaching quality
Electron gun	$3.12^{a,b}$ (0.82)
Beam deflection	$3.27^{a,c}$ (0.95)
Penetration depth	2.73^{b} (0.78)
Electron detector	3.72^{c} (0.79)

Note: Means that do not share a superscript differ at $p < .05$ by the LSD post hoc test.

Table 3. Mean academic performance (ANCOVA adjusted for pre-test scores in % of max. score; standard deviations in brackets)

	Scores	
Subtopic	Expert	Novice
Electron gun	66 (36)	49 (36)
Beam deflection	57 (33)	54 (36)
Penetration depth	61 (31)	39 (33)
Electron detector	79 (29)	33 (37)

Expert academic performance. Academic performance by experts on their own subtopics and novice academic performance on instructed subtopics are presented in Table 3. As expected, experts scored higher overall than novices ($F(1, 106) = 25.51$; $p < .001$).

Correlations with novice academic performance. We first calculated the correlations between novice academic performance and the three variables (a) teaching quality of experts, (b) intrinsic motivation of novice students, and (c) prior topic knowledge of novice students. As expected, novice academic performance correlated significantly with high teaching quality (0.13), high intrinsic motivation (0.15), and high prior topic knowledge (0.43). We explored possible reasons for the low correlations between novice performance and both teaching quality and intrinsic motivation. At the level of individual subtopics we found inconsistencies. For the subtopics *electron gun* and *electron detector*, correlations were significant, in accord with the overall results (Table 4). In contrast, novice academic performance on the subtopics *beam deflection* and *penetration depth* were not correlated with teaching quality or intrinsic motivation, but strongly correlated with prior topic knowledge. These findings can be understood by considering the cognitive demand of each subtopic and its relation to novice academic performance. This dependence on subtopic cognitive demand is addressed in the subsequent section (Research goal 2).

Table 4. Correlations (Pearson's coefficients) with novice academic performance

Subtopic	Teaching quality	Intrinsic motivation of novice students	Prior topic knowledge of novice students
Electron gun	0.30*	0.22*	0.26*
Beam deflection	0.15	0.06	0.63*
Penetration depth	−0.04	0.07	0.50*
Electron detector	0.23*	0.31*	0.38*

Note: $^{*}p \leq .05$.

Research Goal 2: Relationship between Cognitive Demand and Novice Academic Performance

Cognitive demand of the jigsaw tasks. The number of conceptual elements per unit of meaning as proposed by each physics education specialist is depicted in Table 5. As explicated in the Methods section, we used these values to estimate the cognitive demand of each subtopic. Cognitive demand differed significantly between subtopics ($F(3, 22) = 8.77$; $p < .01$)[6] and pair-wise least significant difference (LSD) post hoc analysis revealed that the subtopics *electron gun* and *electron detector* were less cognitively demanding than the subtopics *beam deflection* and *penetration depth*.

The difference in the mean number of conceptual elements per unit of meaning is consistent with the conclusions of Farrington (2011) that working memory is overloaded when processing more than a few elements at a time.

Regression analyses. The relation between teaching quality and novice academic performance depended on the subtopic (Table 4). To explore why teaching quality was associated with better novice academic performance for some subtopics (*electron gun* and *electron detector*) but not others (*beam deflection* and *penetration depth*), we examined the aggregate scores of the less demanding subtopics *electron gun* and *electron detector* and the more demanding subtopics *beam deflection* and *penetration depth* separately. We used teaching quality, intrinsic motivation, and prior topic knowledge of

Table 5. Mean number of conceptual elements (standard deviations in brackets)

Subtopic	Conceptual elements
Electron gun	4.19^{a} (0.62)
Beam deflection	6.08^{b} (0.65)
Penetration depth	5.64^{b} (1.1)
Electron detector	3.97^{a} (0.68)

Note: Means that do not share a superscript differ at $p < .05$ by the LSD post hoc test.

Table 6. Hierarchical linear regression analysis used to predict novice academic performance

	Less demanding subtopics		More demanding subtopics	
Predictor	Non-standardised coefficient (standard error)	*p*	Non-standardised coefficient (standard error)	*p*
Fixed effects				
Intercept	0.39 (0.03)	.00	0.46 (0.03)	.00
Level 1				
Prior topic knowledge	0.44 (0.14)	.00	0.88 (0.13)	.00
Intrinsic motivation of novice students	0.09 (0.03)	.00	−0.01 (0.00)	.81
Level 2				
Teaching quality	0.11 (0.01)	.03	0.00 (0.07)	.97

novice students as independent variables, while novice academic performance served as the dependent variable.

We ran a hierarchical linear regression where teaching quality was included at the teaching group level (level 2), while intrinsic motivation and prior topic knowledge were dealt with on an individual level (level 1). Table 6 shows the non-standardised coefficients and standard errors (in brackets).[7]

For less demanding subtopics, intrinsic motivation of novice students, quality of expert teaching, and prior topic knowledge were significant predictors of novice academic performance (post-test scores), whereas for more demanding subtopics, quality of expert teaching and novice intrinsic motivation both lost their impact on novice academic performance. For more cognitively demanding subtopics, prior topic knowledge was the only significant predictor.

Discussion

As expected, we found a significant, albeit weak, correlation between expert teaching quality and novice post-test scores in the jigsaw classroom study session on the SEM. While this result is consistent with a meta-analysis of teaching effectiveness in conventional classroom settings conducted by Seidel and Shavelson (2007), our result is by no means self-evident. Expert students' explanations in the jigsaw classroom must first be understood by novice students and then trigger deep elaboration of the subject matter to be effective (Webb & Farivar, 1999).

When the cognitive demands of the individual subtopics were included in the statistical analysis, we found that the positive effect of expert teaching quality on novice academic performance was dependent on subtopic being taught. In fact, there was no significant correlation between teaching quality and novice academic performance for the two subtopics with particularly high cognitive demand as assessed by unit of meaning-based concept mapping. The measurement of cognitive demand suggested that two of four subtopics (*beam deflection* and *penetration depth*) were significantly more demanding than the other two subtopics, *electron gun* and *electron detector*.

From the perspective of the CLT, highly interactive elements constituting one concept map must be processed simultaneously in working memory for comprehension of the corresponding unit of meaning. If the number of conceptual elements is too high, cognitive overload can occur, which reduces learning by diverting memory resources from information processing.

To clarify the impact of task cognitive demand on novice learning, we conducted two separate regression analyses. Both used teaching quality, intrinsic motivation of novice students, and prior topic knowledge of novice studies as predictors, and novice academic performance as the dependent variable. We found that the influence of teaching quality and novice intrinsic motivation depended on whether the subtopic was of high or low cognitive demand. When cognitively demanding subtopics were taught, teaching quality and intrinsic motivation of novices lost their impact on novice academic performance. For subtopics of high and low cognitive demand, however, prior topic knowledge was a significant predictor of novice academic performance.

Why did teaching quality have no significant impact on novice academic performance when more cognitively demanding subject matter was taught? We propose two plausible interpretations.

According to the relationship of content knowledge with pedagogical content knowledge, it is reasonable to suppose that students in expert groups who could not acquire adequate knowledge of the subtopic due to the high cognitive demand would perform poorly as teachers (Krauss et al., 2008). For all subtopics, however, the expert students' academic performance was satisfactory and greater than that of the novices on the same subtopic. Moreover, teaching quality was not systematically related to the cognitive demand of the subtopics (Table 2). Teaching quality was ranked in the mid-range and the variance of teaching quality scores was roughly the same for all subtopics. No floor or ceiling effect was present. Teaching quality ranged from high to low, but even when teaching quality was high, novice academic performance was not improved on tests of cognitively demanding material.

Hence, we propose the following alternative interpretation. Cognitively demanding tasks caused high intrinsic cognitive load on novice students due to the higher mean number of conceptual elements per unit of meaning, leading to a severe reduction in germane load and thereby hindering novice learning. Hence, even high-quality expert teaching, characterised by adequate pacing and coherent presentation, failed to support novice learning. Under these circumstances, success depended mainly on prior topic knowledge. Even a high level of intrinsic motivation could not compensate for inadequate prior topic knowledge.

To assess the cognitive demand of each subtopic, we used unit of meaning-based concept mapping as depicted in the Instruments and Measures section. According to CLT, estimation of cognitive demand must account for prior knowledge. This complex process was performed by physics education specialists who simultaneously evaluate both task demand and student expertise on a high inferential level. Which conditions should be met to apply this approach appropriately? The challenge is to put oneself in a typical students' position to structure complex tasks into units of meaning and appropriate elements, and to sequence these in time. The specialists were required to follow one essential guideline; the elements within a unit of meaning must be strongly related, in contrast, the elements of distinct units of meaning are far less interdependent. Two factors are crucial for the success of this method. First, the specialists must be thoroughly familiar with the tasks, and so were requested to work through them in great detail prior to the rating. From studying the material (including performing the experiments and running the simulation), the specialists gained insights into the various levels of abstraction that characterise each task element. For example, a deeper level understanding of the release of electrons off the hot wire is facilitated by the question 'Why must the wire be heated to release electrons?' on the experts' worksheet. Hence, the students are prompted to integrate their prior knowledge on electron thermal motion into the present discussion. Second, the specialists must be as familiar as possible with the students' relevant prior knowledge. According to CLT, this is necessary because the number of conceptual elements depends strongly on prior knowledge. For example, the specialist who developed the units of meaning concept map shown in Figure 2 decided to integrate all processes of electron release from the wire into a single element, the schema 'thermal motion', probably based on his knowledge that details of thermionic emission had been discussed by the teachers in the physics lessons prior to the study (cf. section 'The Learning Unit'). Of course, this rating is still ultimately subjective. However, the physics education specialists were aware of the curriculum (12th grade physics) and of typical procedures for teaching standard topics (e.g. the electron gun). As a result, the expert raters achieved acceptable consensus (cf. Instruments and Measures section).

There are several limitations of this study. Due to the design, no causality can be inferred. In part due to the nature of the tasks and the cooperative learning setting, cognitive load and instructional quality are complex constructs. Hence, as discussed earlier, the operationalisation of the constructs is challenging and the interpretation of the results requires due caution.

The present study was performed in a learning environment that was highly structured, with comprehensive and standardised material given to students in the expert groups, and involved students with a relatively uniform prior topic knowledge base. In more open learning environments, prior topic knowledge is not as clearly defined and more heterogeneous across students. In such a setting, estimation of cognitive demand would be more difficult because cognitive demand strongly depends on prior topic knowledge.

Although expert teaching quality does contribute to novice learning in the jigsaw teaching group, many other factors that may also impact learning were not controlled. As pointed out in the Background section, teaching by explicit explanation is not the only way students learn. Help is often provided in the form of implicit instructional messages because peers are likely to be less skilled than adults in providing explicit explanations (Gillies, 2003). For example, from the context novices may correctly infer that the function of the coil is to generate the magnetic field for the electron beam deflection. Hence, novices may understand the coils' function without experts' explicit explanation. Furthermore, co-construction of knowledge through discussions between all students might foster learning. The quality of these discussions, in turn, depends on factors such as social cohesion. In other words, students will help each other because they have formed bonds within the group (Slavin, 1996).

Important open questions concern the role of experiments and computer simulations in determining cognitive demand. For example, is the subtopic *beam deflection* cognitively demanding due to the high number of conceptual elements per unit of meaning or to the involvement of an experiment that caused additional cognitive load imposed by the split-attention effect? Is the inclusion of experiments and computer simulations detrimental for learning or did they consolidate learning through visualisation and modelling? Is the combination of low relevance of prior topic knowledge and lack of visualisation by an experiment or a computer simulation (cf. subtopic *electron detector* in Table 1) particularly unfavourable, leading to an extra-large performance gap between experts and novices (cf. Table 3)? Clearly, additional studies on the complex interplay between elements in the science learning environments (texts, media, and experiments) are needed. For example, a detailed analysis of the role of experiments and computer simulations according to the proposal of Tesch (2005) would be helpful.

In summary, the main goal of the present study was to analyse the impact of expert teaching quality on the academic performance of instructed novice students in the jigsaw classroom. While teaching quality did enhance performance, further analysis indicated that the benefits of high-quality teaching were reduced for subtopics with high cognitive demand. These results underscore the importance of content selection on the success or failure of cooperative learning with resource interdependence. A practical implication of this study is that cooperative learning may be more successful for less challenging topics. If the tasks are too difficult, the cooperative methods may not foster feelings of competence, a prominent benefit of this teaching model over conventional classroom teaching (Hänze & Berger, 2007).

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (grant numbers HA 3509/1-2 and BE 2652/1-2).

Notes

1. The term 'conceptual element' is used in the cognitive load theory. Throughout the present paper, we employ the term in a broad sense to capture multiple aspects. Apart from scientific concepts, it includes objects, models, phenomena, or principles.
2. That is the dependence of electron kinetic energy on temperature and the need for sufficiently high kinetic energy to overcome the attractive forces within the wire.

3. Retrieved May 22, 2013, from http://www.matter.org.uk/tem/electron_scattering.htm
4. Age was not recorded. In Germany, grade 12 students are typically aged 17–19 years.
5. This item is rated when comprehension difficulties arise.
6. Two values > 10 with distances from the nearest quartile greater than 1.5 times the interquartile range were excluded from analysis.
7. According to Snijders and Bosker (2012, p. 109), for hierarchical linear models the concept of explained proportion of variance (R^2) is problematic. Hence, we refrain from reporting values.

References

Alexander, P. A., Kulikowich, J. M., & Schulze, S. K. (1994). How subject-matter knowledge affects recall and interest. *American Educational Research Journal*, *31*, 313–337.

Aronson, E. (2002). Building empathy, compassion, and achievement in the jigsaw classroom. In J. Aronson (Ed.), *Improving academic achievement. Impact of psychological factors on education* (pp. 209–225). San Diego, CA: Academic Press.

Aronson, E., Blaney, N., Stephan, C., Sikes, J., & Snapp, M. (1978). *The jigsaw classroom*. Beverly Hills, CA: Sage.

Berger, R., & Hänze, M. (2009). Comparison of two small group learning methods in 12th grade physics classes focusing on intrinsic motivation and academic performance. *International Journal of Science Education*, *31*, 1511–1527.

Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook I: The cognitive domain*. New York: David McKay.

Brophy, J. E., & Good, T. L. (1986). Teacher behaviour and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). New York: Macmillan.

Brünken, R., Seufert, T., & Paas, F. (2010). Measuring cognitive load. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 181–202). New York: Cambridge University Press.

Buchs, C., Butera, F., & Mugny, G. (2004). Resource interdependence, student interactions and performance in cooperative learning. *Educational Psychology*, *24*, 291–314.

Chandler, P., & Sweller, J. (1996). Cognitive load while learning to use a computer program. *Applied Cognitive Psychology*, *10*, 151–170.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, *25*, 471–533.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*, 309–319.

Cohen, E. G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research*, *64*, 1–35.

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*, 227–268.

Dreyfus, A. (1986). Manipulating and diversifying the levels of difficulty and task-sophistication of one and the same laboratory exercise. *European Journal of Science Education*, *8*, 17–25.

Edwards, J., & Dall'Alba, G. (1981). Development of a scale of cognitive demand for analysis of printed secondary science materials. *Research in Science Education*, *11*, 158–170.

Fantuzzo, J. W., King, J. A., & Heller, L. R. (1992). Effects of reciprocal peer tutoring on mathematics and school adjustment: A component analysis. *Journal of Educational Psychology*, *84*, 331–339.

Farrington, J. (2011). From the research: Myths worth dispelling. Seven plus or minus two. *Performance Improvement Quarterly*, *23*, 113–116.

Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C. L., Dutka, S., & Katzaroff, M. (1996). The relation between student ability and the quality and effectiveness of explanations. *American Educational Research Journal*, *33*, 631–664.

Galili, I. (1995). Mechanics background influences students' conceptions in electromagnetism. *International Journal of Science Education*, *17*, 371–387.

Gillies, R. M. (2003). Structuring cooperative group work in classrooms. *International Journal of Educational Research*, *39*, 35–49.

Gopal, H., Kleinsmidt, J., & Case, J. (2004). An investigation of tertiary students' understanding of evaporation, condensation and vapour pressure. *International Journal of Science Education*, *26*, 1597–1620.

Hänze, M., & Berger, R. (2007). Cooperative learning, motivational effects and student characteristics: An experimental study comparing cooperative learning and direct instruction in 12th grade physics classes. *Learning and Instruction*, *17*, 29–41.

Hattie, J. (2003, February 24). *Teachers make a difference: What is the research evidence*? Paper presented to Australian Council for Educational Research Annual Conference, Melbourne, 19–21 October. Retrieved from https://www.det.nsw.edu.au/proflearn/docs/pdf/qt_hattie.pdf

Herweg, C., Seidel, T., & Dalehefte, I. M. (2005). Coding manual—clear and structured teaching. In T. Seidel, M. Prenzel, & M. Kobarg (Eds.), *How to run a video study: Technical report of the IPN video study* (pp. 145–164). Münster: Waxmann.

Howe, A. C., & Durr, B. P. (1982). Analysis of an instructional unit for level of cognitive demand. *Journal of Research in Science Teaching*, *19*, 217–224.

Ing, M. (2008). Using instructional sensitivity and instructional opportunities to interpret students' mathematics performance. *Journal of Educational Research and Policy Studies*, *8*, 23–43.

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational Researcher*, *38*, 365–379.

Krauss, S., Brunner, M., Kunter, M., Baumert, J., Blum, W., Neubrand, M., & Jordan, A. (2008). Pedagogical content knowledge and content knowledge of secondary mathematics teachers. *Journal of Educational Psychology*, *100*, 716–725.

Moreno, R. (2009). Constructing knowledge with an agent-based instructional program: A comparison of cooperative and individual meaning making. *Learning and Instruction*, *19*, 433–444.

Neumann, K., Kauertz, A., & Fischer, H. E. (2012). Quality of instruction in science education. In B. J. Fraser, K.G. Tobin, & C. J. McRobbie (Eds.), *Second international handbook of science education* (pp. 247–258). New York: Springer.

Novak, J. D. (1998). *Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, *1*, 117–175.

Renkl, A. (1995). Learning for later teaching: An exploration of mediational links between teaching expectancy and learning results. *Learning and Instruction*, *5*, 21–36.

Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from worked-out examples: The effects of example variability and elicited self-explanations. *Contemporary Educational Psychology*, *23*, 90–108.

Roscoe, R. D., & Chi, M. T. H. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research*, *77*, 534–574.

Rosenshine, B., & Meister, C. (1994). Reciprocal teaching: A review of research. *Review of Educational Research*, *64*, 479–530.

Rosenshine, B., & Stevens, R. (1986). Teaching functions. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 376–391). New York: Macmillan.

Saglam, M., & Millar, R. (2006). Upper high school students' understanding of electromagnetism. *International Journal of Science Education*, *28*, 543–566.

Schiefele, U. (1991). Interest, learning, and motivation. *Educational Psychologist*, *26*, 299–323.

Schnotz, W., & Kürschner, C. (2007). A reconsideration of cognitive load theory. *Educational Psychology Review*, *19*, 469–508.

Seidel, T., Prenzel, M., Rimmele, R., Herweg, C., Kobarg, M., Schwindt, K., & Dalehefte, I.-M. (2007). Science teaching and learning in German physics classrooms. In M. Prenzel (Ed.), *Studies on the educational quality of schools* (pp. 79–99). Münster: Waxmann.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, *77*, 454–499.

Slavin, R. E. (1983). *Cooperative learning*. New York: Longman.

Slavin, R. E. (1996). Research on cooperative learning and achievement: What we know, what we need to know. *Contemporary Educational Psychology*, *21*, 43–69.

Slavin, R. E. (1997). *Educational psychology*. Boston, MA: Allyn and Bacon.

Slavin, R. E., Hurley, E. A., & Chamberlain, A. (2003). Cooperative learning and achievement: Theory and research. In W. M. Reynolds, G. E. Miller, & I. B. Weiner (Eds.), *Handbook of psychology, educational psychology* (Vol. 7, pp. 177–197). New York: Wiley.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis. An introduction to basic and advanced multilevel modeling* (2nd ed.). Los Angeles: Sage.

Souvignier, E., & Kronenberger, J. (2007). Cooperative learning in third graders' jigsaw groups for mathematics and science with and without questioning training. *British Journal of Educational Psychology*, *77*, 755–771.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*, 257–285.

Sweller, J. (2010). Cognitive load theory: Recent theoretical advances. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 29–47). New York: Cambridge University Press.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. New York: Springer.

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction*, *12*, 185–233.

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*, 251–296.

Tesch, M. (2005). Coding manual—experiments in physics lessons. In T. Seidel, M. Prenzel, & M. Kobarg (Eds.), *How to run a video study: Technical report of the IPN video study* (pp. 91–107). Münster: Waxmann.

Webb, N. M. (1991). Task-related verbal interaction and mathematics learning in small groups. *Journal for Research in Mathematics Education*, *22*, 366–389.

Webb, N. M., & Farivar, S. (1999). Developing productive group interaction in middle school mathematics. In A. M. O'Donnell, & A. King (Eds.), *Cognitive perspectives on peer learning* (pp. 117–149). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Weinert, F. E., Schrader, F.-W., & Helmke, A. (1989). Quality of instruction and achievement outcomes. *International Journal of Educational Research*, *13*, 895–914.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*, 68–81.

Wittwer, J., & Renkl, A. (2008). Why instructional explanations often do not work: A framework for understanding the effectiveness of instructional explanations. *Educational Psychologist*, *43*, 49–64.

Appendix 1

Subtopic 1: Electron Gun

Electrons are released from a hot metallic wire due to thermionic emission. Subsequently, they are accelerated by a strong electric field. The principle of the electron gun (including thermionic emission and acceleration voltage) was taught in another context (the television tube) in prior to the study. Both concepts encompass additional conceptual elements. To understand thermionic emission within the framework of the atomic model, students need to know that increased temperature means greater kinetic energy of electrons. This enhances the fraction of electrons with sufficient high kinetic energy to leave the wire. To understand acceleration of the electrons, students must know that charged particles experience forces in electric fields and that (net) forces lead to acceleration.

From research on students' understanding of evaporation, it is known that even advanced students generally do not realise that kinetic energy of particles is inhomogeneously distributed (Gopal, Kleinsmidt, & Case, 2004). By analogy, this idea is relevant for understanding thermionic emission on a microscopic level.

Many students predict the trajectory of a charged particle (e.g. an electron) along a field line (Galili, 1995) because the concepts of velocity and acceleration are not well differentiated. Since the precise trajectories of electrons are not crucial for a basic understanding of how the electron gun works, this aspect is not discussed.

Subtopic 2: Deflection of the Electron Beam

The electron beam released from the electron gun is deflected by the magnetic field of a coil. The strength of the magnetic field is controlled by the electric current in the coil. By cyclical changes of the electric current, the electron beam can be scanned over a sample systematically.

The goal is to understand the connection between the direction of the current in the coil and the direction of the beam deflection. In order to understand this interrelation, students first have to realise that electrons in the coil move to the positive pole of the electric supply. Students must learn to predict the direction of the magnetic field based on the 'left-hand rule'. If the direction of the magnetic field is known, students determine the Lorentz force on the electron beam according to the 'three-finger rule'. Both rules encompass several linked conceptual elements. The rules have been introduced in previous lessons, but in a different context.

Understanding the motion of electrons in a magnetic field may be hindered by adopting inappropriate analogies (Saglam & Millar, 2006). Thus, students do not adequately differentiate between electric and magnetic fields. Furthermore, students may think of magnetic fields as a 'flow' that pushes the entering charge.

Subtopic 3: Penetration Depth of Electrons in the Sample

The high-energy electrons of the beam penetrate the sample and interact with atoms. The path of the electrons is determined mainly by attraction to the atomic nuclei. The penetration depth depends on two factors. First, the higher the atomic number (the number of protons in the nucleus), the stronger the deflection of the electron and the shorter the penetration depth. Second, the smaller the primary electron energy, the shorter the penetration depth.

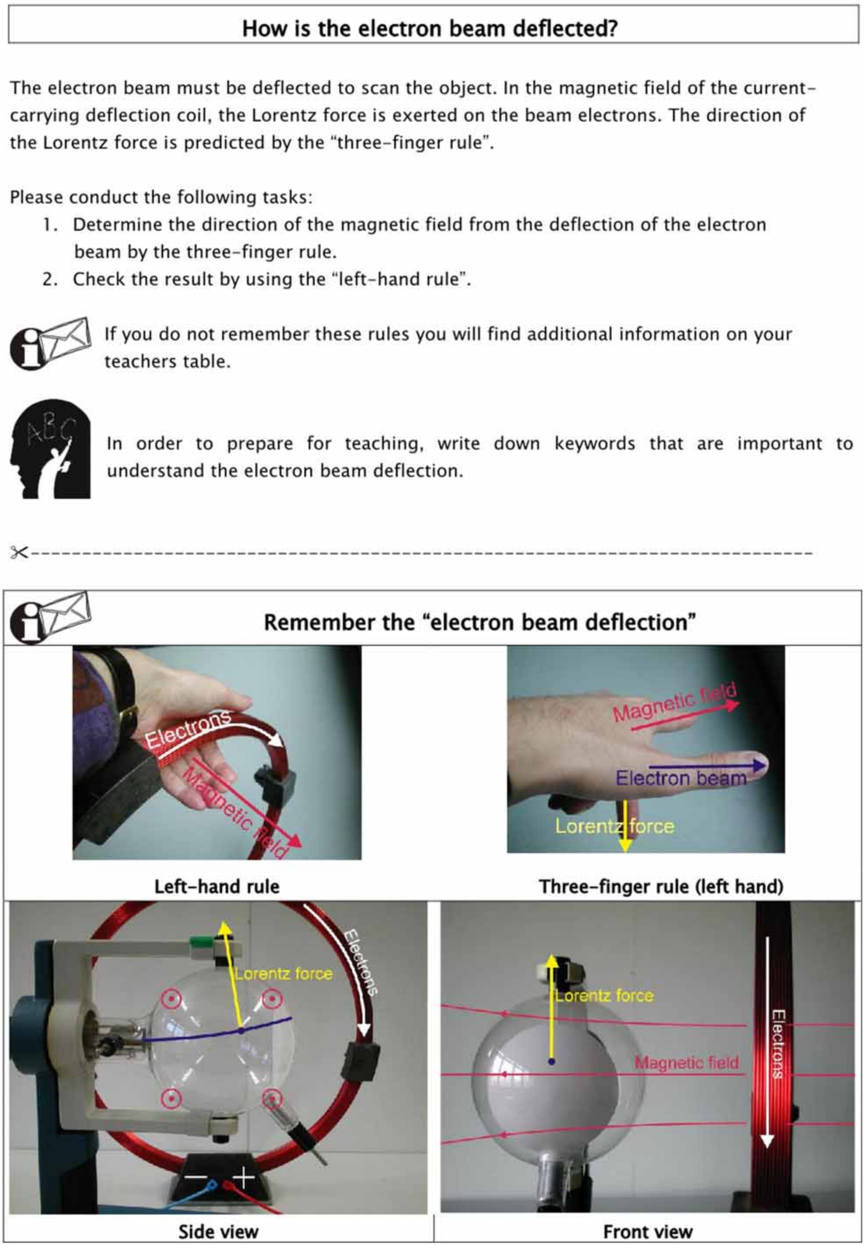
Students are requested to vary both atomic number and electron energy independently and observe the change in the electron paths on a micrometer scale. These observations have to be interpreted in light of the electron–nucleus interaction (i.e. on a nanometre scale). Both effects (of energy and atomic number) are independent. Nevertheless, students must recognise that one variable must be held constant to assess the effect of the other. Furthermore, both lines of reasoning involve several steps. Apart from the electrostatic interaction of charges, no further prior topic knowledge is required to accomplish the task.

Subtopic 4: Electron Detector

After release from the sample surface, the electrons enter the 'Everhart-Thornley detector'. First, they are accelerated by a strong electric field in order to excite a scintillator. The scintillator emits light that is transmitted through a glass fibre to a metal surface that releases electrons (photoelectric effect). These electrons are accelerated by strong electric fields onto a series of electrodes in an electron multiplier. Through impact ionisation, an avalanche of free electrons is generated that controls the pixel-by-pixel brightness of the monitor.

No deeper insights into the microscopic foundations of the scintillation process or the photoelectric effect are intended because these aspects will be discussed in detail in the next school year. Thus, no task-specific prior knowledge is required. The learning goal is to memorise the various steps of the detection process.

Appendix 2: Example of Experts' Texts (Subtopic 'Electron Beam Deflection')



How is the electron beam deflected?

The electron beam must be deflected to scan the object. In the magnetic field of the current-carrying deflection coil, the Lorentz force is exerted on the beam electrons. The direction of the Lorentz force is predicted by the "three-finger rule".

Please conduct the following tasks:

1. Determine the direction of the magnetic field from the deflection of the electron beam by the three-finger rule.
2. Check the result by using the "left-hand rule".

If you do not remember these rules you will find additional information on your teachers table.

In order to prepare for teaching, write down keywords that are important to understand the electron beam deflection.

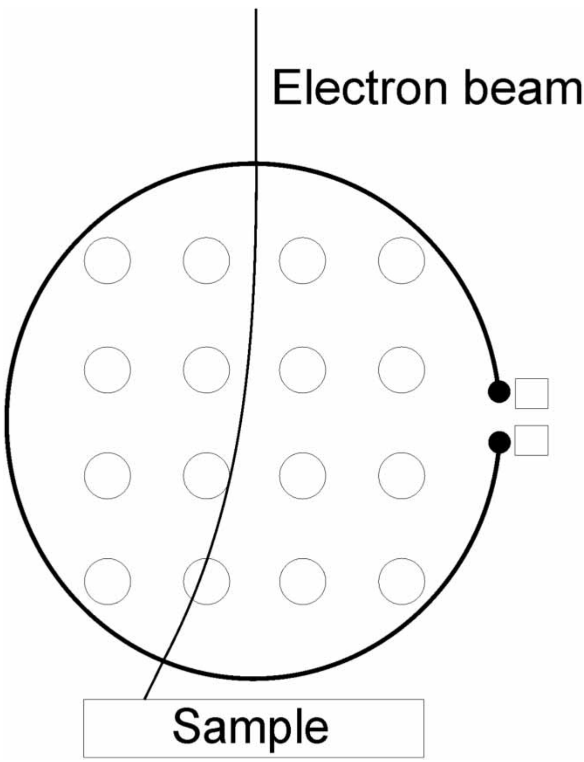
Remember the "electron beam deflection"

Figure A1. Example of an experts' worksheet (upper part) and the corresponding written help on the teachers' desk (lower part).

Appendix 3: Post-test

1. Electron Gun (3 points)
Please explain, by means of a drawing, how electrons in the electron gun of an electron microscope are released and accelerated.

2. Deflection of the Electron Beam (5 points)
Electrons enter the magnetic field of the deflection coil according to the drawing below.
a) Sketch the direction of the magnetic field lines on the drawing (field lines out of the plane of drawing: ⊙, into the plane of drawing: $\times$). Please give a detailed explanation for your choice based on the relevant rules.
b) Sketch the poles (+/−) of the voltage generator into the drawing (☐). Please give a detailed explanation for your choice based on the relevant rules.



3. Penetration Depth of Electrons in the Sample (6 points)
a) Why is the length of a primary electrons' trajectory always greater than the penetration depth? Please give an explanation at the atomic level.

b) How does the penetration depth of the primary electrons depend on the main mechanisms that we discussed in school? For each mechanism, give an explanation based on the scattering by the atomic nuclei.

4. Electron Detector (5 points)
Please explain, by means of a drawing, how the secondary electrons are detected.

Scoring scheme

1. Electron Gun (3 points)
Current to the heating voltage heats the wire. (1 P.) The enhanced temperature is related to an enhanced electrons' velocity. Electrons with sufficiently high temperature leave the wire. (1 P.) The free electrons (negative charged) are accelerated by the positive pole of the acceleration voltage. (1 P.)

2. Deflection of the Electron Beam (5 points)
a) Adequate drawing and explanation ("three-finger rule") (3 P.)
b) Adequate drawing and explanation ("left-hand rule") (2 P.)
Note: Consistent drawing without explanation: 1 P.

3. Penetration Depth of Electrons in the Sample (6 points)
a) The primary electrons are deflected by the electric field of the sample atoms leading to a "zig-zag" trajectory. Hence, the penetration depth is smaller than the length of the trajectory. (2 P.)
b) 1. For a given primary electrons' kinetic energy, the higher the atomic number, the lower the penetration depth (1 P.) because more protons lead to a stronger deflection of the primary electrons due to the stronger electric field of the nucleus (1 P.)
2. For a given atomic number, the higher the primary electrons' kinetic energy, the higher the penetration depth (1 P.), because the interaction time between nucleus and primary electron is smaller. (1 P.)

4. Electron Detector (5 points)
Electrons entering the detector are accelerated by a strong electric field. (1 P.) Subsequently, the electrons strike the scintillator material. (1 P.) The emitted light is transmitted through a glass fibre to a metal surface/photocathode (1 P.), causing the release of electrons (photoelectric effect). (1 P.) These electrons are accelerated by strong electric fields onto a series of electrodes in an electron multiplier. Through impact ionisation, an avalanche of free electrons is generated that controls the pixel-by-pixel brightness of the monitor. (1 P.)