Fostering the Acquisition of Scientific Reasoning with Video Modeling Examples and Inquiry Tasks

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

M.Sc. Juliane Maren Kant

aus Malsch

Tübingen

2016

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	26.01.2017
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Katharina Scheiter
2. Berichterstatter:	Jun.-Prof. Dr. Kerstin Oschatz

Acknowledgements

As individuals, we may start out with the curiosity and disposition to be little scientists, but it is a long journey from information seeking to skilled scientific reasoning, with the help of many scaffolds along the way.

Morris, Croker, Masnick, and Zimmerman (2012, p. 74)

I want to thank all the "scaffolds" along the way of my dissertation:

- My first supervisor Katharina Scheiter for giving me freedom in what to do research about and how to do it, for the best text feedback I could have wished for, for encouragement in my toughest times, for being a role model in doing politics, and for very interesting AG lunches!
- My second supervisor Kerstin Oschatz for having good ideas, asking critical questions, providing emotional support, and preventing me from being unemployed!
- My colleagues Katrin Schleinschok, Sina Müller, Christina Artemenko, Michèle Suhlmann, Leona Hellwig, Fabian Lang, Juliane Richter, and Tzu-Ling Hua for helping me in data collection, proofreading, giving me advice in every condition of life, having lunch together, having chats about research and the world, making music together and helping me in overcoming my timidity to talk in English!
- My IT support André Klemke and Markus Überall for teaching me programming basics, for solving every IT problem, for help in preparing the hardware equipment and for being reachable during data collection, when I was most nervous!
- My "models" Anke Hoffmann and Mark Sinzger for repeating the same or similar texts over and over again and keeping concentrated until the video modeling examples were good!
- My research assistants Annika Thierfelder, Isabella Geiger, Anika Staib, Kathrin Raab, Julia Kollmer and my friend Dorothea Dalig for watching and coding all the students' experiments and helping in data collection at schools!
- Mareike Bierlich for being the heart of the LEAD Graduate School & Research Network, for being always reachable and for solving small and bigger problems during my dissertation!
- Hanna Giselbrecht for being my best friend, for sharing all the secrets, and for supporting me at virtually any time when I needed her even from the other side of the world!

- The principles Dr. Peter Gilbert, Uwe Müller, and Jürgen Mittag as well as the teachers Benjamin Bücking, Peter Lürßen, Frank Hönig, Natalie Weber, Jochen Ebert, and Andreas Weber for their interest in my study, for giving me the opportunity to test my training program at their schools and for their patience and help during data collection!
- All the students who participated in my studies for learning patiently and sometimes even enthusiastically with my training program!
- My parents Margit Jost-Kant and Hubert Kant, and my brother Paul Kant for getting me in contact with schools, for encouraging me, and for supporting my decisions!
- My beloved colleague and partner Thomas Lösch for all the little and big things you have done for me, for all the good ideas concerning my research, for your endless support, for believing in me, for your enthusiasm for research, for helping in data collection and for doing trips and journeys, and all the other things apart from research!

Contents

Introduction

To participate in our knowledge-based societies and to make sense of the vast amount of scientific knowledge provided by printed and digital media, individuals need *scientific reasoning* skills (Fischer et al., 2014; Trilling & Fadel, 2009). Scientific reasoning includes the skills needed to understand the scientific process across disciplines, to evaluate the validity of scientific claims, to assess the relevance of scientific results, and to apply scientific concepts and methods in order to generate new knowledge. Scientific reasoning thereby can be understood as a thinking style that forms an essential skill for dealing with scientific issues within modern everyday life. Thus, scientific reasoning is a component of scientific literacy, which is an important aim of general education (Bybee, 1997; KMK, 2004). Consequently, the acquisition of scientific reasoning is considered a main goal of science education (National Research Council, 2012; OECD, 2007).

In general, there are two promising approaches to promoting the acquisition of scientific reasoning skills in schools. The first approach is *inquiry learning*, which advocates for learning by doing. In inquiry learning, students learn scientific reasoning by engaging in scientific reasoning activities. In the natural sciences, for example, students develop questions, make observations, design experiments, and collect, analyze, and interpret data to investigate a phenomenon (Lazonder & Harmsen, 2016). The second approach is *example-based learning*, a form of direct instruction. Example-based learning uses pre-structured examples to explicitly teach certain skills (van Gog & Rummel, 2010). To learn scientific reasoning skills, for instance, students study examples showing how to develop questions, make observations, design experiments, and collect, analyze, and interpret data to investigate a phenomenon. Thus, whereas in inquiry learning students act like scientists (learning by doing), in example-based learning students are shown how to act like scientists (direct instruction). There has been a long-standing debate about the relative effectiveness of the two teaching philosophies.

Inquiry learning entails authentic and information-rich settings that offer the possibility to teach students about the complex nature of scientific reasoning (Chinn & Malhotra, 2002; Kirschner, Sweller, & Clark, 2006). When students, for instance, conduct experiments about chemical reactions or about density in physics with sinking and floating objects they must carefully plan their experiments and might be confronted with measurement problems. However, engaging in inquiry activities is difficult for students, especially when they have low prior knowledge. Since inquiry environments are often very complex, this involves the danger

of students simply playing with the materials without learning the underlying reasoning principles. Thus, learners need instructional guidance in inquiry learning (Alfieri, Brooks, Aldrich, & Tenenbaum, 2011; Lazonder & Harmsen, 2016).

Example-based learning, in contrast, is per se highly structured and guided. Examples offer a step-by-step expert solution for a given problem. Thus, students can easily follow and internalize the solution procedure. Studying examples can especially help students with low prior knowledge to acquire new cognitive skills such as scientific reasoning. However, examples only benefit learning if students process them deeply. Just reading an example involves the danger of giving students the impression that they have understood everything when they have not. Consequently, they might terminate studying before they have learned everything in the example.

Since both approaches are associated with benefits and drawbacks, the present thesis investigated how to foster students' acquisition of scientific reasoning skills at schools with inquiry and example-based learning. For this purpose, I developed a digital training program that uses inquiry tasks with virtual experiments and video modeling examples showing how to conduct virtual experiments. In a first step, I examined whether there are benefits of combining the two approaches over learning from just one approach. A combination of both approaches furthermore raised the question of how to sequence inquiry and example-based learning activities. In a second step, I used a beneficial combination of both approaches. Because there are often multiple learning goals in schools, for example acquiring skills and knowledge, I examined how to design the examples in the combined approach to foster the acquisition of scientific reasoning skills and domain knowledge simultaneously.

The present thesis is structured into three main parts. First, Section 1 provides the theoretical background, including an introduction of the main concepts *scientific reasoning*, *inquiry learning* and *example-based learning*, and derives the thesis's research questions. Second, Sections 2 and 3 describe the two experiments building the core of this thesis. Experiment 1 targeted the delivery of examples and investigated how the sequence of video modeling examples and inquiry tasks influenced students' acquisition of scientific reasoning skills in a classroom setting. Results showed a relatively clear benefit of providing learners with video modeling examples before or instead of engaging them in inquiry tasks. Therefore, experiment 2 targeted the design of examples. It investigated how video modeling examples can be optimized to foster the acquisition of scientific reasoning skills and domain knowledge

simultaneously. Finally, Section 4 discusses the results of the two experiments and their implications on how to foster students' acquisition of scientific reasoning skills at schools with inquiry and example-based learning.

1 Theory

The first part of the present thesis outlines the definition, measurement, and development of scientific reasoning. Next, two promising approaches to foster the acquisition of scientific reasoning in schools are introduced. First, inquiry learning with physical and virtual experiments is explicated before addressing the effectiveness of inquiry learning. Second, example-based learning is introduced as an alternative to inquiry learning. Then, factors influencing the effectiveness of example-based learning with a focus on the delivery and the design of the examples are discussed. Finally, the research questions of the two experiments of the present thesis addressing delivery and design of examples in inquiry learning are presented.

1.1 Scientific reasoning: Definition, measurement and development

Scientific reasoning has been investigated by different research disciplines such as developmental psychology (e.g., Inhelder & Piaget, 1958; Kuhn, 2010), cognitive and educational psychology (e.g., Klahr & Dunbar, 1988), and research on science education (e.g., Osborne, 2013). These research disciplines have used various terminologies such as scientific thinking, scientific discovery or inquiry skills as well as different definitions for scientific reasoning (Fischer et al., 2014; Klahr & Dunbar, 1988; Kuhn, 2010; Kuhn & Franklin, 2006; Morris, Croker, Masnick, & Zimmerman, 2012; C. Zimmerman, 2000, 2007). However, all definitions converge on the notion that scientific reasoning includes generating hypotheses, testing hypotheses with experiments, and evaluating evidence of the experiments with regard to the hypotheses. The following paragraphs introduce the most common conceptualizations of scientific reasoning ranging from general (scientific reasoning as intentional knowledge seeking) to more specific definitions (scientific reasoning as encoding information and using strategies). Finally, the different definitions are integrated.

1.1.1 Definition of scientific reasoning

At the most general level, scientific reasoning can be defined as an intentional knowledge-seeking process with the goal of coordinating theory and evidence (Kuhn, 2010; Kuhn & Franklin, 2006). It can be conceptualized as a way of thinking and acquiring knowledge about the world (Kuhn, 2010). According to Kuhn (2010), this intentional knowledge-seeking process, like scientific investigations, have four major phases: inquiry, analysis, inference, and argument. In the inquiry phase, the goals of the investigation and the research questions must be generated. A researcher could, for instance, want to determine which factors help to predict the risk of earthquakes (Kuhn & Pease, 2008). A research question pertaining to this goal might

be: Does the soil type (igneous or sedimentary) influence the earthquake risk? A prerequisite for generating goals and research questions is to acknowledge that one's own existing knowledge is incomplete. Recognizing knowledge gaps can result in the intention to seek new knowledge, for example, in the form of evidence. In the analysis phase, a skilled scientific reasoner will thus access data to collect evidence for his or her research question. In our example, the researcher could collect data of similar regions with different soil types along with their earthquake risk. In the inference phase, this evidence has to be coordinated with a theory resulting in either congruence or discrepancy (Kuhn, 2010). If the new evidence fits an existing theory, it will be incorporated into existing knowledge. If the new evidence contradicts an existing theory, a skilled scientific reasoner will revise his or her theory to be compatible with the evidence (Kuhn, 2010). In our example, the researcher could analyze the data to investigate whether one soil type is always associated with a higher earthquake risk. The argument phase, finally, makes scientific reasoning social and extends it into real life thinking. In this phase, claims are debated with other people (Kuhn, 2010). Thus, our researcher could travel to a conference and exchange his results with other researchers investigating earthquake risk. The complete knowledge-seeking process is guided by meta-level skills. Procedural processes involve the selection, application and monitoring of knowledge-seeking strategies, whereas declarative understanding involves epistemic beliefs about knowledge and knowing in science (Kuhn, 2010).

Recently, a similarly broad definition has been proposed as a starting point for interdisciplinary research (Fischer et al., 2014). According to this definition, scientific reasoning and argumentation

> include the knowledge and skills involved in different epistemic activities (problem identification, questioning, hypothesis generation, construction of artefacts, evidence generation, evidence evaluation, drawing conclusions as well as communicating and scrutinising scientific reasoning and its results) in the context of three different epistemic modes (advancing theory building about natural and social phenomena, science-based reasoning in practice, and artefact-centred scientific reasoning). (Fischer et al., 2014, p. 39)

Again, the three components of generating hypotheses and evidence, and evaluating this evidence are included in this definition. Additionally and similar to Kuhn (2010), argumentation as well as epistemic activities and modes are emphasized. Epistemic modes are used to describe

the motivations behind scientific reasoning and argumentation. The first epistemic mode (advancing theory building about natural and social phenomena) is mainly concerned with understanding science, the second epistemic mode (science-based reasoning in practice) is mainly concerned with the use of science, and the third epistemic mode combines understanding and use (artefact-centered scientific reasoning). It could be argued that the first epistemic mode might only be relevant for professional researchers and the second epistemic mode only for practitioners. However, it can also be argued that all epistemic modes are relevant for students considering that new knowledge or new theories for students do not have to be new for the world. In addition, Fischer et al. (2014) suppose that the epistemic activities and modes might be relevant for all scientific domains. However, they emphasize that domains differ with regard to the nature and the weight of different epistemic activities. Thus, according to this definition, scientific reasoning involves domain-general as well as domain-specific aspects (Fischer et al., 2014).

Klahr and Dunbar (1988) detail the skills that are involved in scientific reasoning. According to their *Scientific Discovery as Dual Search (SDDS)* model, scientific reasoning can be seen as a search process that resembles a problem solving task (Klahr & Dunbar, 1988). Thus, the starting point for a scientific discovery is a complex problem that one would like to solve. Participants in the studies of Klahr and Dunbar (1988), for example, were taught how to use a computer-controlled robot tank and then asked to discover how an unknown key (the RPT key) of the robot worked. Additionally, participants were told that the repeat key could take a numerical value (*N*). The exact function of the repeat key, however, had to be discovered by participants. SDDS proposes that scientific reasoning requires search within and between two related problem spaces: the hypothesis space and the experiment space. The hypothesis space consists of all hypotheses that an individual creates during the discovery process. An example for a hypothesis in the studies by Klahr and Dunbar (1988) is: The unknown key of the robot will repeat the last *N* instructions. The experiment space consists of all possible experiments that could be conducted. Experiments to discover how the unknown key worked consisted of all possible programs that could be written for the robot that include the unknown key. According to SDDS, the scientific discovery is controlled by three major skills: (1) searching hypothesis space, (2) testing hypothesis and (3) evaluating evidence (Klahr & Dunbar, 1988). (1) The aim of searching the hypothesis space is to generate a fully specified hypothesis. This involves two sub-processes. First, a kind of schema containing several possibly relevant variables must be generated. The schema can be generated either by searching prior knowledge or by inducing it from the results of previous experiments. For example, in the experiments by

Klahr and Dunbar (1988) the complete schema contained the variables role of *N*, type of element to be repeated, the boundaries of the repeated element, and the number of repetitions. In a second step, specific values were assigned to the variables in the schema to build a fully specified hypothesis. Again, either prior knowledge or previous experimental outcomes can be used for assigning variable values. A person with programming experience, for example, might assume that the role of *N* might be a counter, that is, it might indicate the number of repetitions. The resulting fully specified hypothesis can then be tested. (2) Testing a hypothesis consists of three sub-processes: designing an experiment, making a prediction and running the experiment. First, participants in the studies by Klahr and Dunbar (1988) designed an experiment. For this purpose, they had to determine a focal variable, that is, they decided which independent variable of a hypothesis would be tested. Next, they set a value for this specific variable. Additionally, all other variables had to be set to values to define the complete experiment. For example, a participant could decide to test if the role of *N* could be a counter and set *N* to three. Second, the current hypothesis and the current experiment are used to make a specific prediction about the results of the experiment. For example, the participant could predict that the robot will repeat an action three times. Third, the experiment is run, results are observed, and compared to the expectations. In the example above, the participant would check whether the robot repeated the action three times. (3) Finally, the evidence must be evaluated by reviewing the results and deciding, whether the evidence is sufficient to accept or reject the hypothesis. If there is not enough evidence, hypothesis testing begins again (Klahr & Dunbar, 1988). The SDDS model was able to explain the different experimenting behaviors of participants in the studies by Klahr and Dunbar (1988). Overall, the SDDS model describes the skills that are common in all definitions of scientific reasoning. In addition, it indicates a similarity between scientific reasoning and problem solving.

The similarity between scientific reasoning and problem solving is also addressed in a definition by Zimmerman and colleagues (Morris et al., 2012; C. Zimmerman, 2000, 2007). C. Zimmerman (2007, p. 173) defines scientific reasoning "as the application of the methods or principles of scientific inquiry to reasoning or problem-solving situations". Moreover, the cognitive mechanisms and metacognitive processes underlying the common scientific reasoning skills (generating and testing hypotheses and evaluating evidence) are explicated. Important cognitive mechanisms are encoding and using strategies (Morris et al., 2012). During encoding, a mental model of information is created in memory, that is, information that we direct attention to is represented (Siegler, 1989). Encoding is essential for any kind of reasoning since information has to be represented in memory before it can be used to reason (Morris et

al., 2012). A second cognitive mechanism is the use of strategies. A strategy contains a sequence of actions that leads from an initial state to a goal state (Morris et al., 2012). There are strategies for all skills involved in scientific reasoning. For example, one strategy is concerned with the *generation of hypotheses* (de Jong & van Joolingen, 1998). Hypotheses should describe the relationship between two variables in a way that can be falsified with an experiment. Thus, the variables as well as the relationship should be measurable. Another very prominent strategy in the context of scientific reasoning is the *control-of-variables strategy* (*CVS*; Chen & Klahr, 1999). It states that all variables except for the one being tested should be held constant across experimental trials to yield conclusive results. The CVS is considered "a basic, domain-general strategy that allows valid inferences and is an important strategic acquisition because it constraints the search of possible experiments" (C. Zimmerman, 2007). Since the CVS is fundamental for drawing conclusions about causal relationships in science, there is a vast amount of research on how to foster this strategy (for a meta-analysis see Schwichow, Croker, Zimmerman, Höffler, & Härtig, 2016). In addition, for successful development and usage of these strategies, metacognitive processes are necessary. It is, for example, important to know when and why a certain scientific reasoning strategy should be applied (Morris et al., 2012) or what constitutes valid scientific knowledge (i.e. epistemic beliefs about the nature of science; Kuhn, 2010).

The following paragraph offers an integration of the central aspects of the above-mentioned definitions of scientific reasoning. Scientific reasoning can be defined as intentional knowledge-seeking process with the aim of coordinating theory and evidence (Kuhn, 2010) in order to solve a problem (Klahr & Dunbar, 1988; C. Zimmerman, 2000). This process requires search in two different problem spaces: hypothesis space and experiment space (Klahr & Dunbar, 1988). The main skills that guide this search are: (1) generating hypotheses, (2) testing hypotheses, and (3) evaluating evidence. Underlying these skills are cognitive mechanisms such as encoding information and using strategies (e.g., the control-of-variables strategy) and metacognitive processes such as epistemic beliefs (Morris et al., 2012). Some definitions also include as an important aspect the ability to communicate results to others, or argumentation (Fischer et al., 2014; Kuhn, 2010). However, I will focus on the three main scientific reasoning skills (generating hypotheses, testing hypotheses with experiments, evaluating evidence of experiments with regard to the hypotheses) of Klahr and Dunbar (1988) that are common to all definitions. These scientific reasoning skills will be used to organize the empirical studies of the next sections.

1.1.2 Measurement of scientific reasoning

To develop a training program to foster scientific reasoning, it is important to know how to operationalize or assess the construct of interest (Bortz, 1984). Otherwise, it is not possible to describe changes or judge the effectiveness of any program. There are at least three possibilities to assess the scientific reasoning skills: assessment with achievement tests with multiple-choice items (e.g., Pant et al., 2013), assessment with physical materials (e.g., Siegler & Chen, 1998), and assessment with virtual simulated materials (e.g., Gobert, Sao Pedro, Raziuddin, & Baker, 2013). The following paragraphs describe these three approaches to evaluating the three main scientific reasoning skills: generating hypotheses, conducting experiments, and evaluating evidence.

A simple and economic approach to assess scientific reasoning can be operationalized through achievement tests with multiple-choice items. In this case, learners answer questions or solve tasks. To correctly answer the respective questions or solve the respective tasks, it is assumed that scientific reasoning skills are required. An example for this can be found in the IQB National Assessment Study 2012 (Pant et al., 2013). Scientific reasoning was assessed as a sub-area of the content area scientific inquiry. Several aspects of scientific reasoning such as formulation of a question, hypothesis, study design, and data evaluation were assessed by asking Grade 9 students to solve written tasks. To assess the aspect ‘hypothesis’, for example, students had to infer from a text and a table describing an experiment which hypothesis was tested with this experiment (Wellnitz, Fischer, Kauertz, Neumann, & Pant, 2012). To assess ‘study design’ students had to read a short text about a study with a research question and several variables. Subsequently, they were asked to decide which variables to change and which to keep constant to answer the research question (Köller, 2008). Finally, to assess ‘data evaluation’ students read a short text describing an experiment and were asked which of four provided conclusions would be valid (Köller, 2008). There are many examples of achievement tests that assess scientific reasoning skills (Blair, 1940; Chang et al., 2011; Glug, 2009; Hardy, Kleickmann, & Koerber, 2010; Hartmann, Upmeier zu Belzen, Krüger, & Pant, 2015; Klos, Henke, Kieren, Walpuski, & Sumfleth, 2008; Koenen, 2014; Lawson, 1978, 2009; Shahali & Halim, 2010; Wellnitz et al., 2012). However, a possible disadvantage of achievement tests is that they might not be valid indicators of actual scientific reasoning behavior. That is, they might not be predictive of scientific reasoning in a real-world context. Thus, achievement tests entail the risk of assessing inert knowledge, that is, knowledge that can be reproduced in assessment situations but that would not be spontaneously applied to real life problem-solving situations (Renkl, Mandl, & Gruber, 1996). In addition, there has been criticism as to whether

achievement tests are well suited to assess scientific reasoning skills or whether they rather capture rote understanding of science (DeBoer, Abell, Regan, & Wilson, 2008; Gobert et al., 2013; Quellmalz et al., 2013).

Assessment with physical materials provides an alternative to multiple-choice achievement tests. In this approach, learners observe or conduct relatively real investigations. Their experimenting behavior is observed and coded. To assess hypothesis generation, Piekny and Maehler (2013), for example, presented cards with fantasy animals to children. The animals were presented in groups of families with characteristic body parts. After each card the children hypothesized which body parts were important in order to belong to a certain family (Piekny & Maehler, 2013). The number of hypotheses that were in line with the presented evidence was used as a measure of the ability to generate hypotheses (Piekny & Maehler, 2013). Alternatively, Siegler and Chen (1998) asked children to predict which side of a balance scale with several weights would go down if two wooden blocks placed under the arms of the scale were removed. The predictions were then classified according to underlying rules and were used as a measure for generating hypotheses (Siegler & Chen, 1998). To assess the ability to design experiments, learners can be asked to actually design physical experiments. Chen and Klahr (1999), for example, used three physical experiments (spring task, ramp task, and sinking task) to assess the ability to design controlled experiments. In the ramp task, children designed experiments to decide how four different variables (steepness of the ramp, surface of the ramp, starting gate, kind of ball) affected the distance that a ball rolls. The number of valid comparisons the children made (i.e., pairs of trials in which only one independent variable was varied while all other independent variables were kept constant) was used as a measure for the ability to design experiments (Chen & Klahr, 1999). In a study by Siegler and Liebert (1975), students had to detect how to move an electric train by discovering a specific combination of four on/off switches. However, the train was actually controlled by a secret switch so that the discovery of the solution could be postponed until all 16 possible combinations had been generated (Siegler & Liebert, 1975). In this study, the ability to design experiments was assessed via the number of generated combinations. Finally, the ability to evaluate evidence can be assessed by asking participants to interpret experimental outcomes. For example, learners often have to interpret covariation data in relation to competing hypotheses (Piekny & Maehler, 2013). Two studies with young children as participants used pictures of children with either a red or a green chewing gum and bad or healthy teeth. Participants saw sets of pictures either indicating perfect covariation between chewing gum color and health of teeth, imperfect covariation or non-covariation. Participants then had to decide if the evidence supported a

causal relationship between color of chewing gum and health of teeth. The number of correct answers was used as a measure for the ability to evaluate evidence (Koerber, Sodian, Thoermer, & Nett, 2005; Piekny & Maehler, 2013). Although this method can produce a more realistic picture of scientific reasoning ability, assessing with physical materials is often time-consuming, materials can be expensive, and there are unobservable phenomena like chemical reactions that cannot be captured with this approach (de Jong, Linn, & Zacharia, 2013). In addition, it is often not possible to test several participants in one class simultaneously with physical experiments (Linn, 2000).

Therefore, it can be beneficial to use virtual materials in the form of virtual simulated experiments to assess scientific reasoning skills (Gobert et al., 2013; National Research Council, 2001; Quellmalz, Timms, & Schneider, 2009). To assess the ability to generate hypotheses, van Joolingen and de Jong (1993), for example, analyzed the log files of students working in an inquiry environment. The number of hypotheses stated by students was counted. In addition, every hypothesis was evaluated in terms of correctness, precision and domain of applicability. The ability to design experiments with virtual materials is often assessed through the correct application of the control-of-variables strategy. Gobert et al. (2013), for example, assessed students' ability to design controlled experiments using log files and educational data mining techniques. Features analyzed were, for example, a count of variable changes, the number of pairwise repeated trials, or the number of pairwise controlled trials (Gobert et al., 2013). Other researchers have used the number of unique simulation experiments (i.e., experiments that have not been previously run with the same values) or the application of the control-of-variables strategy as measures (e.g., Mulder, Lazonder, & de Jong, 2014). Finally, the ability to evaluate evidence can be assessed by checking whether learners draw correct conclusions from their experiments. In a study by Kuhn and Dean (2005), for example, students investigated if different binary variables (e.g., soil type) had an influence on earthquake risks in a simulation-based inquiry environment. In the end, participants indicated which of the variables they thought made a difference in the earthquake risk. The number of valid inferences was used as a measure for the ability to evaluate evidence and a valid inference was defined as a determinate inference that was supported by evidence generated by the students (Kuhn & Dean, 2005). Virtual experiments can provide a relatively authentic environment for scientific reasoning (Chinn & Malhotra, 2002). Thus, this approach is relatively close to real experimenting behavior. In addition, virtual experiments require less time and costs for schools. Finally, process data from log files of experiments can give insights into the learning process of learners (Gobert et al., 2013; Pedro, Gobert, & Baker, 2012). Hence, simulated experiments

are a promising approach to an economic and behavior-based assessment of scientific reasoning. Adequate assessment of scientific reasoning skills, in turn, is an important prerequisite to map the development of these skills.

1.1.3 Development of scientific reasoning

Research on the development of scientific reasoning dates back to Inhelder and Piaget (1958). They investigated the development of cognitive abilities from childhood to adolescence. According to Inhelder and Piaget (1958), children are only able to reason scientifically when they reach the final stage of cognitive development, that is, the stage of formal operations. In this stage, children should become able to reason about their reasoning. As a consequence, other reasoning abilities like the systematic combination and isolation of variables, proportional, and correlational reasoning should emerge (Kuhn & Franklin, 2006). However, research in the past decades has shown that children are capable to reason scientifically earlier than expected (Koerber, Sodian, Kropf, Mayer, & Schwippert, 2011; Koerber et al., 2005; Kuhn, 2010; Kuhn & Franklin, 2006; Wilkening & Sodian, 2005). In the following, research on the development of the abilities to generate hypotheses, design experiments, and evaluate evidence will briefly be reviewed alongside problems that learners may encounter.

The ability to generate hypotheses does not emerge until the beginning of elementary school. Preschoolers have difficulties formulating hypotheses (Piekny & Maehler, 2013). However, young elementary schools students are able to distinguish between testing a hypothesis and producing specific result (Sodian, Zaitchik, & Carey, 1991). From elementary school onward children are able to generate hypotheses, but tend to generate hypotheses that are in line with their prior beliefs (Lazonder & Harmsen, 2016; C. Zimmerman, 2007). In addition, especially younger children often tend to focus on a single plausible hypothesis (Klahr, Fay, & Dunbar, 1993). However, considering many alternative hypotheses has been shown to lead to more successful experimentation (Klahr & Dunbar, 1988), yet, teenagers and even adults generate very few hypotheses spontaneously (Njoo & de Jong, 1993). Finally, learners are often not able to adapt their hypothesis based on the data they collect. Learners tend to keep their hypotheses even in the presence of disconfirming evidence (Klahr & Dunbar, 1988).

Designing experiments to generate or test hypotheses seems to be difficult for all age groups (Lazonder & Harmsen, 2016). Whereas five-year-olds are not yet able to distinguish between testing a hypothesis and producing a results, six-year-olds begin to develop this ability

(Piekny, Grube, & Maehler, 2014; Sodian et al., 1991). Still, especially young learners often conduct experiments without any hypothesis. Instead of testing a hypothesis, they try to produce a specific result (engineering approach; Schauble, Klopfer, & Raghavan, 1991). Students around the age of ten become able to conduct unconfounded experiments investigating the relationship between an independent variable and a dependent variable that clearly covary (Kanari & Millar, 2004). However, when variables do not perfectly covary, students had problems investigating relationships with unconfounded experiments (Kanari & Millar, 2004). Even though the ability to design experiments increases with age, it remains a difficult task for learners. Learners often design inconclusive experiments, that is, they vary several variables in one experimental trial rather than applying the control-of-variables strategy (Glaser, Schauble, Raghavan, & Zeitz, 1992; Keselman, 2003). Consequently, they cannot draw any conclusions from their experimental results. Furthermore, learners exhibit inefficient experimentation behavior. For example, they repeat the same experiment several times or devote experimental time to variables that are already well understood (Klahr et al., 1993).

Preschoolers or young elementary school students are already able to evaluate perfect covariation data (Koerber et al., 2011, 2005; Piekny et al., 2014; Piekny & Maehler, 2013). The ability to evaluate imperfect covariation data seems to be more demanding (Inhelder & Piaget, 1958; Kuhn & Phelps, 1982). This skill develops slowly and hardly ever reaches maturity (Lazonder & Harmsen, 2016). Instead, learners fail in drawing the right conclusions from their experimental results. That is, learners frequently infer that a variable is causal when indeed it is not and make inferences that are consistent with their prior beliefs or based on a single instance of covariation (C. Zimmerman, 2007).

In conclusion, even if many precursors of scientific reasoning already develop during childhood, skilled scientific reasoning does not develop routinely (Kuhn & Franklin, 2006). Instead, the developmental trajectory of scientific reasoning is slow and requires instructional support (Morris et al., 2012).

1.2 Fostering the acquisition of scientific reasoning

A debate remains about how to best support the development of scientific reasoning skills (Hmelo-Silver, Duncan, & Chinn, 2007; Kirschner et al., 2006; Klahr & Nigam, 2005; Kuhn & Dean, 2005; R. E. Mayer, 2004). On the one hand, there is the constructivist idea that scientific reasoning and science content knowledge can best be learned through inquiry or discovery learning, where learners have to discover scientific phenomena and construct new knowledge on their own (Hmelo-Silver et al., 2007; Kuhn & Dean, 2005). On the other hand, there is the information-processing approach advocating direct or explicit instruction such as example-based learning as being more appropriate to foster scientific reasoning (Kirschner et al., 2006; Klahr & Nigam, 2005; R. E. Mayer, 2004). Thus, whereas inquiry learning argues for learning by doing or problem-solving (Pedaste et al., 2015), example-based learning argues for learning by being told. In the following, both approaches are described.

1.2.1 Inquiry learning

According to proponents of the inquiry learning approach, scientific reasoning (as well as science knowledge; see below) can best be learned through reasoning scientifically, that is, through learning by doing. Thus, in inquiry learning, learners "follow methods and practices similar to those of professional scientists in order to construct knowledge" (Pedaste et al., 2015, p. 48). Through engagement in these methods and practices, learners should develop the necessary scientific reasoning skills. What exactly constitutes inquiry learning? Up until now there is no consistent and generally accepted definition of inquiry learning (Klahr & Nigam, 2005). There seems to be a consensus that inquiry learning involves self-directed investigations by learners in order to solve a complex problem (Lazonder & Harmsen, 2016). According to Alfieri et al. (2011, p. 2) inquiry learning "occurs whenever the learner is not provided with the target information or conceptual understanding and must find it independently and with only the provided materials". A recent meta-analysis further specified the definition of inquiry learning in line with the standards of the National Research Council (2012). In inquiry learning

> students conduct experiments, make observations or collect information in order to infer the principles underlying a topic or domain. These investigations are governed by one or more research questions, either provided by the teacher or proposed by the students; adhere (loosely) to the stages outlined in the scientific method; and can be performed with computer simulations, virtual labs, tangible materials, or existing databases. (Lazonder & Harmsen, 2016, p. 2).

This definition of inquiry learning shares similarities with definitions of scientific reasoning. This is a result of the proposition of the constructivist approach that scientific reasoning can best be learned through reasoning scientifically, that is through performing inquiry activities. In this sense, inquiry learning is means and ends simultaneously. Bruner (1961, p. 7), for example, has argued that inquiry learning could foster "the art and technique of inquiry" itself. However, inquiry learning can also be used as a means for learning conceptual domain knowledge in domains such as biology, physics and chemistry. In the following, I describe different ways of implementing inquiry learning in schools.

1.2.1.1 Implementation of inquiry learning: Physical and virtual experiments

Inquiry learning has been applied in schools since the discovery learning movement in the 1960s (Lazonder & Harmsen, 2016). Whereas in the beginning it was mainly used to teach science content, inquiry learning has since also been used to teach science process skills (Lazonder & Harmsen, 2016).

One way to implement inquiry learning in schools is to ask learners to conduct physical experiments with tangible materials. Physical experiments offer the possibility for students to acquire hands-on laboratory skills. Moreover, students can gain an adequate picture of the complexity of science including unexpected events such as measurement error (de Jong et al., 2013). Finally, the tactile information that learners get during experimenting with tangible materials can foster the understanding of science concepts, according to research on embodied cognition. For example, experiencing torque enhances students' understanding of angular momentum, compared to observing another person experiencing torque (Kontra, Lyons, Fischer, & Beilock, 2015). However, especially in schools, experiments with tangible materials also have certain disadvantages. They are relatively time-consuming and sometimes require expensive or even dangerous materials (e.g., toxic chemicals). Additionally, physical experiments may require learners to handle a large number of information elements simultaneously which may overwhelm students' limited cognitive resources (Sweller, van Merriënboer, & Paas, 1998; Tuovinen & Sweller, 1999). An alternative, which might circumvent these disadvantages, are virtual experiments with computer simulations.

Virtual experiments have the advantage that "reality can be adapted" (de Jong et al., 2013, p. 305). Thus, the complexity of scientific phenomena can be reduced, for example, by removing confusing details or by highlighting important aspects of an experiment (Trundle & Bell, 2010). This might help to reduce the amount of information that learners have to process

simultaneously. Furthermore, virtual experiments provide an opportunity for students to conduct experiments on otherwise unobservable phenomena like planetary movements (de Jong et al., 2013). Virtual experiments also require less setup time than physical experiments and results can be obtained in shorter time frames (Zacharia, Olympiou, & Papaevripidou, 2008).

Importantly, studies comparing the effectiveness of learning with physical and virtual experiments so far have found no performance differences regarding learners' conceptual understanding and inquiry skills (for a review see de Jong et al., 2013). One study, for example, engaged children in creating and testing mousetrap cars (Klahr, Triona, & Williams, 2007). The aim was to discover which features would make a car that travels as far as possible. Children worked either with physical or virtual cars. There were no differences between the groups in their knowledge about causal factors and in their ability to design respective cars (Klahr et al., 2007). Another study compared the effectiveness of teaching the control-of-variables strategy with physical or virtual springs and weights (Triona & Klahr, 2003). The two types of materials were equally effective in instructing children how to design unconfounded experiments (Triona & Klahr, 2003). Consequently, virtual simulated experiments offer a promising alternative to physical experiments for inquiry learning in schools. However, the question remains whether inquiry learning is effective in enhancing scientific reasoning skills and domain knowledge compared to more expository forms of instruction.

1.2.1.2 Effectiveness of inquiry learning

Why could inquiry learning be effective? Advocators of inquiry learning have argued that all learning involves the construction of new knowledge and thus is a constructivist process (Hmelo-Silver et al., 2007). Since inquiry learning requires learners to construct their own solutions it might be better suited for meaningful constructivist learning than more expository forms of instruction (Kirschner et al., 2006). Moreover, "involving students in activities that demand inquiry as a means to fostering inquiry skills" (Dean & Kuhn, 2007, p. 386), that is, learning by doing is intuitively plausible. Another potential advantage of inquiry learning might be that it often entails authentic and information-rich settings (Kirschner et al., 2006). This offers the possibility to teach students about the complex nature of scientific reasoning and enhance their epistemological understanding (Chinn & Malhotra, 2002). Finally, it has been argued that inquiry learning might be beneficial for long-term learning and transfer. Dean and Kuhn (2007), for example, compared the performance of three different groups that learned the control-of-variables strategy over an extended period of several weeks. Whereas an inquiry group engaged in computer-based problems that required the CVS for effective solution over

12 sessions, a direct instruction group received a single session designed to teach CVS. Finally, a direct instruction plus inquiry group received a combination of the first two groups. A post-instruction assessment immediately after the direct instruction session showed an advantage of the two direct instruction groups over the inquiry group. However, the posttest and transfer performance after ten weeks and after 17 weeks was higher for the inquiry groups than for the group receiving only direct instruction (Dean & Kuhn, 2007). In this study, inquiry activities fostered long-term learning and transfer.

However, since its inception in the 1960s inquiry learning has also received plenty of criticism. Because inquiry learning is often labor-intensive and time-inefficient, there might not be enough time to discover every topic of the science curriculum. In addition, direct instruction methods have shown to be highly effective, especially for complex procedures that learners are unlikely to discover on their own (e.g., Anderson, Corbett, Koedinger, & Pelletier, 1995). Contrary to the results of the study by Dean and Kuhn (2007), for example, Klahr and Nigam (2005) showed that direct instruction was more effective than inquiry learning in teaching children the control-of-variables strategy. More children acquired mastery of the CVS through direct instruction than through inquiry learning. However, there were also children in the inquiry learning condition that mastered CVS. In addition, all children who mastered CVS could transfer what they had learned to evaluating science poster fairs (Klahr & Nigam, 2005). The strongest argument against inquiry learning, however, is that inquiry learning might exceed human working memory limitations (see Kirschner et al., 2006 for a review). Human working memory can only process a limited amount of information simultaneously (Miller, 1956). Since authentic inquiry environments can be very complex, learners must process a large amount of information in addition to the relevant contents or skills. Thus, learners might be overwhelmed because not enough resources remain for meaningful learning (Kirschner et al., 2006; Tuovinen & Sweller, 1999). Reviewing the literature of the past decades, R. E. Mayer (2004) has argued that inquiry learning with no or minimal guidance should be abandoned given the lack of studies showing that inquiry learning improves learning outcomes.

So, whose claim has the stronger empirical support? Is inquiry learning more effective than direct instruction or the other way around? This question was addressed in a meta-analysis by Alfieri et al. (2011). The authors examined the effects of inquiry learning versus direct instruction over 108 studies. The mean effect size of all studies was $d = -0.38$, indicating that inquiry learning with no or minimal guidance was less beneficial for learning than direct instruction. Thus, research speaks clearly against inquiry learning, but contemporary inquiry-

based methods might be effective for learning nevertheless since they include extensive guidance for learners (Hmelo-Silver et al., 2007).

Guidance can be defined "as any form of assistance offered before and/or during the inquiry learning process that aims to simplify, provide a view on, elicit, supplant, or prescribe the scientific reasoning skills involved" (Lazonder & Harmsen, 2016, p. 7). Thus, guidance in this sense helps learners regarding single aspects in the problem-solving process. Learners, however, still must solve the problems on their own. There is a large body of research on inquiry learning that includes guidance (Collins, Brown, & Newman, 1989; Davis, 2000; Guzdial, 1994; Jackson, Stratford, Krajcik, & Soloway, 1994; Reiser, 2004; Toth, Suthers, & Lesgold, 2002). Consequently, Alfieri et al. (2011) conducted a second meta-analysis comparing the effect of guided inquiry learning with other forms of instruction (including unguided inquiry learning and direct instruction). The mean effect size of 65 studies was $d = 0.30$, indicating an advantage of guided inquiry over other forms of instruction (Alfieri et al., 2011). The type of other instruction did not moderate the findings. That is, guided inquiry led to better learning outcomes than direct teaching, providing explanations, unguided inquiry or baseline activities. Only the effectiveness of one form of direct instruction, namely worked examples, was not different from guided inquiry (Alfieri et al., 2011).

This result is further supported by a recent meta-analysis by Lazonder and Harmsen (2016). They investigated the effectiveness of different types of guidance on learning activities, performance success, and learning outcomes in inquiry learning, respectively. For this purpose, they used the typology of guidance in inquiry learning proposed by de Jong and Lazonder (2014). This typology is organized according to the specificity of the guidance learners need to successfully perform inquiry. Guidance in inquiry learning can be offered as *process constraints*, *status overviews*, *prompts*, *heuristics*, *scaffolds* or *explanations* (de Jong & Lazonder, 2014). Process constraints are the least specific method for providing guidance to learners. Process constraints break the inquiry task down into several subtasks that are manageable for learners. Status overviews show the learners what they have already performed and/or how well they performed. Thus, they make the task progress visible. Prompts are cues that remind the learner to perform a certain action. They can either be given by a teacher or embedded in the learning environment (Lazonder & Harmsen, 2016). Heuristics are cues similar to prompts but with additional information about how to perform the prompted action. Scaffolds explain or take over the demanding parts of an action. When the learners' skills increase, they are usually faded out. Finally, explanations are the most specific type of guidance.

They specify exactly how to perform a particular action. Results of the meta-analysis by Lazonder and Harmsen (2016) including 72 studies showed that overall guidance had a significant positive influence on learning activities, performance success, and learning outcomes. Moreover, the effect of guidance on performance success was moderated by the specificity of guidance. That is, learners perform better during an inquiry (e.g., generate more valid inferences, better concept maps, or complete more assignments correctly) when supported by more specific types of guidance such as scaffolds or explanations (Lazonder & Harmsen, 2016).

In conclusion, pure inquiry learning with minimal or no guidance is less effective for learning than direct instruction (Alfieri et al., 2011; Kirschner et al., 2006). Guided inquiry learning, in contrast, including scaffolding or feedback or requiring learners to generate answers to experimenters' questions or explain aspects of the task to themselves, can lead to better learning outcomes than direct instruction or unguided inquiry. However, the effectiveness of worked examples, one specific form of direct instruction, was not different from guided inquiry. Thus, in the following, example-based learning is introduced as an alternative approach to inquiry learning for fostering scientific reasoning skills.

1.2.2 Example-based learning

Advocators of example-based learning argue "that it would be impossible (not to mention quite dangerous) for a human being to discover by one's one experience the vast amounts of knowledge that our ancestors developed over thousands of years" (van Gog & Rummel, 2010, pp. 155–156). It appears to be much more efficient to borrow knowledge from others through learning by observing (Paas & Sweller, 2012; van Gog & Rummel, 2010; van Merriënboer & Sweller, 2005). In example-based learning, therefore, examples are used to show how a specific problem may be solved. Thus, in an example, the problem-solving process is explained. Learners are expected to comprehend the problem-solving process without solving the problem on their own. Examples can consist either of *worked examples* that provide a written account of how a problem should be solved or *modeling examples* in which a model (e.g., teacher or peer learner) demonstrates how to solve a problem (van Gog & Rummel, 2010). In both cases, an example contains an underlying abstract principle and a surface story or problem context in which the principle is explained.

1.2.2.1 Effectiveness of example-based learning

Research on worked examples has consistently shown that it is beneficial for novice learners to study worked examples containing a step-by-step expert solution to a problem, rather than solving problems on their own (Atkinson, Derry, Renkl, & Wortham, 2000; Cooper & Sweller, 1987; Renkl, 2014; Sweller & Cooper, 1985). This so called worked-example effect is usually explained in terms of cognitive load theory, that is, the different cognitive processes evoked by studying examples or solving problems (Sweller et al., 1998). When learners with low prior knowledge solve problems they are forced to rely on weak problem-solving strategies such as means-ends analysis (Sweller, 1988). Means-ends analysis requires learners to consider the current problem state, the goal state, and to search for a way to reduce the distance between the two states. This imposes a high working memory load on learners, which is not effective for learning (Sweller, 1988). Thus, learners might not be able to construct a cognitive schema of how such a problem should be solved. Studying worked examples, on the other hand, prevents learners from using weak problem-solving strategies. Instead, learners can use all available working memory capacity to focus their attention on problem states and useful solution steps, and to build a cognitive problem-solving schema (Sweller et al., 1998). Thus, learners reach better learning outcomes with less investment of time and effort.

Traditionally, worked examples have been used to foster performance in highly structured cognitive tasks such as algebra (Cooper & Sweller, 1987; Sweller & Cooper, 1985), statistics (Quilici & Mayer, 1996), geometry (Paas & van Merriënboer, 1994; Schwonke et al., 2009) or physics (Kalyuga, Chandler, Tuovinen, & Sweller, 2001; Reisslein, Atkinson, Seeling, & Reisslein, 2006). Sweller and Cooper (1985), for example, compared groups of students who either studied worked examples or solved analogous problems when learning algebra. Results showed that studying worked examples required less time than solving problems. Additionally, studying worked examples enabled students to solve problems in a posttest more rapidly and with fewer errors (Sweller & Cooper, 1985).

Modeling examples, in contrast, have more often been used to foster less structured cognitive skills such as writing (Braaksma, Rijlaarsdam, & van den Bergh, 2002), assertive communication (Decker, 1980), collaboration (Rummel & Spada, 2005; Rummel, Spada, & Hauser, 2009), and scientific reasoning (Mulder et al., 2014). Mastering such less structured cognitive skills requires often iterative or cyclical processes, which partly depend on preceding steps (Hilbert, Renkl, Kessler, & Reiss, 2008; Mulder et al., 2014). During scientific reasoning, for example, generated evidence must be evaluated with regard to a hypothesis. If the evidence

is not sufficient to reject or support the hypothesis, a skilled scientific reasoner returns to experimenting. Thus, a skilled scientific reasoner has to consider previously performed actions and results to decide what to do next (Mulder et al., 2014). Text-based worked examples, however, which usually present a straightforward algorithmic solution procedure, are not well suited to capture the rationale of this cyclical process. In modeling examples, in contrast, the model can explain his thoughts or heuristics when trying to solve a problem (Hilbert et al., 2008). Consequently, they might be more suited to teach less structured cognitive skills such as scientific reasoning.

Mulder et al. (2014) were the first to use video modeling examples to foster scientific reasoning skills. In their study, high-school students investigated a simulation of an electrical circuit with a modelling tool. Students' inquiry task was to examine and model the influence and interactions of the elements in the electrical circuit. One group of students could additionally consult video modeling examples that explained the required activities and how to perform them, whereas another group did not receive this support. Video modeling examples contained a dynamic screen capture of a person performing an inquiry task and demonstrating scientific reasoning strategies, for example, the control-of-variables strategy (Mulder et al., 2014). Results showed that video modeling examples enhanced students' inquiry behavior and the quality of the models of electric circuits they created compared to students who did not see video modeling examples. However, the quality of students' models was rather modest in both groups and there were no differences in domain knowledge between the groups (Mulder et al., 2014). In their discussion, the authors suggested to optimize the examples to make them more effective.

Hence, it is important to know what makes example-based learning effective - especially when combined with inquiry learning. There a several factors that influence the effectiveness of example-based learning such as *learner characteristics*, *delivery of examples* and *design of examples* (van Gog & Rummel, 2010). These aspects will be explained in the next sections.

1.2.2.2 Learner characteristics

The most important learner characteristic that influences the effectiveness of examples is learners' prior knowledge. As has been described above, example-based learning is especially effective for learners with low prior knowledge (van Gog & Rummel, 2010). Studying examples is beneficial for novices in a domain because it helps them to construct cognitive schemata. However, if learners already possess prior knowledge in the form of schemata,

studying examples can be ineffective or even detrimental for learning. This phenomenon is called the expertise reversal effect (Kalyuga et al., 2001). Whereas learner characteristics can be influenced externally only to a certain extent, delivery and design of examples can more easily be manipulated to optimize the effectiveness of example-based learning. Hence, these two aspects were investigated in Study 1 and 2 of the present thesis, respectively.

1.2.2.3 Delivery of examples

The delivery of examples targets the questions of if and how examples should be combined with problems. According to the worked example effect, studying examples leads to better learning outcomes than solving problems (Renkl, 2014). However, is possible that combining examples and problems might be even more effective than studying examples only. Combining examples and problems might have several advantages over studying examples only. First, there might be motivational advantages. For instance, it might be more motivating for learners to solve a similar problem immediately after example study because it is a more active form of learning than studying another example (Sweller & Cooper, 1985). Second, combining examples and problems allows learners to practice solving problems. This might enable learners to recognize deficiencies in their performance and motivate them to study the next example more closely (van Gog & Rummel, 2010). Practicing solving problems might also prevent inert knowledge, that is, knowledge that can be retrieved in assessment situations but it is not applied to solve problems in real life (Renkl et al., 1996).

In addition, if there are advantages of combining examples and problems over studying examples only, the question arises of how examples and problems should be combined. This question pertains to the timing or the sequence of examples and problems (van Gog & Rummel, 2010). Examples could be presented first followed by problems (example-problem pairs) or the other way around (problem-example pairs). Presenting an example before a problem might be advantageous since studying an example first could reduce cognitive load in learners and help them to build a cognitive problem-solving schema (van Gog, Kester, & Paas, 2011). Solving a problem subsequently might enable learners to stabilize and apply this problem-solving schema.

However, studying an example first could also have detrimental effects on learning since it can give learners an illusion of understanding (Baars, van Gog, de Bruin, & Paas, 2016, 2014; Renkl & Atkinson, 2002). An illusion of understanding means that learners think they have understood everything when they actually have not. Illusions of understanding can thus result

in overconfidence. That is, learners' judgments of their future performance are higher than their actual future performance. During schema acquisition, for example, overconfident learners might terminate studying before a schema is constructed at all or before all relevant elements of a schema are encoded and incorporated. Thus, overconfidence might prevent or impair the acquisition of a problem-solving schema through invalid regulation processes (Dunlosky & Rawson, 2012). Illusions of understanding might even be more likely to occur when using video modeling examples. Dynamic visualizations like videos are commonly associated with entertainment. Therefore, students may underestimate the effort necessary to understand what is being conveyed through a dynamic visualization (underwhelming effect; Lowe, 2004).

Consequently, it might also be beneficial to present problems before examples. Solving a problem first might enable students to recognize deficiencies in their own performance which might direct their attention to those aspects during studying the subsequent example (Hausmann, van de Sande, & VanLehn, 2008; van Gog et al., 2011). Thus, problem solving might prepare students for studying examples. Similarly, solving a problem first can be considered an active generative task (Baars, van Gog, et al., 2014) that gives learners valuable information about their current state of learning. Thus, learners might become aware of what they already have or have not yet learned. This could help them to study a subsequent example with a specific focus on their knowledge gaps.

Research on the sequencing of examples and problems has resulted in mixed evidence. On the one hand, there is research speaking in favor of presenting examples before problems. Two studies in the context of worked example research, for example, found an advantage for presenting examples first. Both studies have investigated the effectiveness of examples only, examples followed by problems (example-problems pairs) and problems followed by examples (problems-example pairs) compared with problems only (Leppink, Paas, van Gog, van der Vleuten, & van Merriënboer, 2014; van Gog et al., 2011). Van Gog et al. (2011) found that participants who learned to troubleshoot electrical circuits via example-problems pairs or examples only indicated lower cognitive load and showed better learning outcomes than participants who learned with problems-example pairs or problems only. Moreover, participants who learned with example-problems pairs did not differ from participants who learned with examples only. Similarly, participants who learned with problems-example pairs did not differ from participants who learned with problems only. Leppink et al. (2014) replicated the advantage of studying an example over solving a problem first in a different

domain (application of Bayes' theorem). Thus, research on worked examples speaks in favor of presenting an example first followed by either a problem or another example.

In addition, research on sequencing instruction and inquiry further corroborated the result that presenting an example before a problem can be beneficial for novices. Several studies underscore a positive effect on learning outcomes of presenting information before inquiry (Barzilai & Blau, 2014; Lazonder, Hagemans, & de Jong, 2010; Wecker et al., 2013). Barzilai and Blau (2014), for example, compared the effectiveness of providing a scaffold including examples before or after an inquiry activity to an inquiry activity without scaffolds. Results showed that learners who studied the scaffold before the inquiry exhibited higher problem-solving performance in a posttest than learners who either studied scaffolds after the inquiry or not at all (Barzilai & Blau, 2014).

On the other hand, there is also research speaking in favor of presenting problems before examples. This sequence has been extensively investigated in research on preparation for future learning (Schwartz & Martin, 2004) and productive failure (Kapur, 2012). A study by Arena and Schwartz (2014), for example, investigated if a videogame could prepare students for future formal instruction. In the video game players needed to infer the shape of probability distributions to perform well which can be considered a problem solving or inquiry task. The formal instruction consisted of a written text including several examples about probability distributions. Results showed that students who first played the game and then read the passage learned more than participants who only read the passage (Arena & Schwartz, 2014).

Advantages of presenting problems before examples have also been shown by research on productive failure. In this approach, students are presented with a problem with a rich database and asked to devise several solutions (Kapur, 2012). Since students get no hints about relevant features for problem solution in the database, they are most often unable to create the canonical solution. The struggle to find solutions is thought to trigger a general awareness of their knowledge gaps and prepare them for the following instruction phase. In the first part of the instruction phase, the teacher demonstrates the limitations of typical students' solutions before modeling the canonical solution (Loibl & Rummel, 2014). This sequence of problem-solving prior to instruction has been shown to result in better conceptual understanding than instruction prior to problem-solving (Kapur, 2012; Loibl & Rummel, 2014). Loibl and Rummel (2014) showed that problem solving prior to instruction indeed triggered a global awareness of

knowledge gaps that was beneficial for learning when combined with instruction with student solutions.

An advantage of presenting problems before examples might also be expected from the perspective of self-regulated learning. Becoming aware of knowledge gaps can be considered a metacognitive process which is an important aspect of self-regulated learning (Nelson & Narens, 1990). Self-regulated learning can be defined as a cyclical process involving feelings, thoughts and actions that are oriented towards attaining a learning goal (B. J. Zimmerman, 2002). A successful self-regulated learner not only sets learning goals and chooses appropriate strategies to achieve the goal (cognitive processes) but also monitors the learning progress and regulates his/her actions accordingly (metacognitive processes; B. J. Zimmerman, 2002). Only when learners are able to accurately monitor their learning process can they regulate their learning process adequately. That is, high monitoring accuracy, which is usually assessed by asking students to predict their future test performance (judgments of learning, JoLs) and relating this judgement to their actual test performance, is associated with higher learning outcomes (Thiede, Anderson, & Therriault, 2003). One way to improve monitoring accuracy is to ask learners to perform an active generation task. Research on learning from expository text has shown that generation activities such as writing keywords or summaries or asking learners to complete diagrams indeed improved the monitoring accuracy of learners (Thiede & Anderson, 2003; Thiede et al., 2003; van Loon, de Bruin, van Gog, van Merriënboer, & Dunlosky, 2014). Solving a problem can also be considered a generative activity. Thus, from the perspective of self-regulated learning, presenting a problem first may result in better learning outcomes than presenting an example first because it leads to a higher monitoring accuracy and a subsequently better regulation.

Overall, the delivery of examples targets the questions of whehter and how examples should be combined with problems. It is yet unclear if there are advantages of combining examples and problems over studying examples only. If there should be advantages of combining examples and problems the question arises of how to sequence them. Studying an example first might reduce cognitive load in learners and help them to build a problem-solving schema. Subsequently solving a problem might help them to stabilize and use the problem-solving schema (Barzilai & Blau, 2014; Lazonder et al., 2010; Leppink et al., 2014; van Gog et al., 2011; Wecker et al., 2013). However, studying an example first might also result in overconfidence and thus lead learners to terminate studying before they have learned everything. In contrast, solving a problem first might make learners aware of knowledge gaps,

that is, enhance their monitoring accuracy and prepare and motivate them to study a subsequent example more closely (Arena & Schwartz, 2014; Loibl & Rummel, 2014; Thiede et al., 2003). Despite the amount of research dedicated to the sequencing of examples and problems, up until now there has been no attempt to investigate effects of sequencing of video modeling examples and inquiry tasks on scientific reasoning skills. This issue will be addressed in Study 1 of the present thesis.

1.2.2.4 Design of examples

Another important aspect that affects the effectiveness of example-based learning is the design of the examples. Examples have to be designed according to certain instructional design principles in order to be effective (Atkinson et al., 2000). Two important design features in the context of example-based learning are the *instructional approach* and the *arrangement* of examples according to the principles they convey and the context in which they are embedded (Renkl, 2014, 2015).

The instructional approach addresses the question of how to present the abstract principle or structural feature of a worked example. As described in Section 1.2.2, a worked example contains an underlying abstract principle and a surface story or problem context in which the principle is explained. The instructional approach describes whether the abstract principle is introduced before or within the problem context. In a *deductive approach,* the abstract principle is introduced first, followed by examples in which the principle is applied (Renkl, 2015; Ross & Kilbane, 1997). In an *inductive approach,* on the other hand, the abstract principle is not introduced explicitly to the learner. Instead, learners receive only the examples in which the principle is embedded. In this approach, learners must induce the principle themselves. Both approaches can be further supported by differential prompts. Learners profit most from the deductive approach if they are prompted to explain the solution to themselves (Renkl, Stark, Gruber, & Mandl, 1998). Such a self-explanation can comprise elaborations on the application conditions and goals of domain principles, or it can comprise relations between solution steps and domain principles (Renkl et al., 1998). On the other hand, studies show that learners profit most from the inductive approach when they are prompted to compare different examples (e.g., Gentner, Loewenstein, & Thompson, 2003).

Moreover, deductive and inductive approaches have differential effects on different knowledge facets. Deductive approaches that present a rule followed by an example foster the acquisition of declarative knowledge and concepts (Seidel, Blomberg, & Renkl, 2013;

Tomlinson & Hunt, 1971). Pre-service teachers, for example, who learned principles for teaching and learning with a deductive approach were better able to reproduce declarative knowledge of the subject compared to pre-service teachers who learned with an inductive approach (Seidel et al., 2013). Inductive approaches, on the other hand, seem to specifically facilitate the acquisition of skills (Gentner et al., 2003; Seidel et al., 2013). In the study of Seidel et al. (2013), pre-service teachers who learned with an inductive approach were better able to apply principles for teaching and learning during lesson planning compared to pre-service-teachers who learned with a deductive approach. In another study, novice learners who learned a negotiation strategy through inductive examples were more likely to apply the strategy in a simulated negotiation than a baseline group without examples (Gentner et al., 2003). To sum up, both the deductive and the inductive instructional approach seem to be effective with regard to different types of learning outcomes.

Whereas the instructional approach, a design feature of examples, targets the question of how to introduce the principle of an example, another design feature addresses the question of how to *arrange* multiple examples regarding their structural and context features. Research on worked examples has shown that learners profit most by studying multiple examples for one principle (Renkl, 2014). According to Quilici and Mayer (1996), one principle can either be taught in a *surface-emphasizing* or in a *structure-emphasizing* way. Thereby, it varies whether one principle is taught with examples using the same or different surface features or contexts. In a surface-emphasizing arrangement, one principle is taught with several examples using the same story context. In such an arrangement, one principle is always associated with the same surface features. Thus, a surface-emphasizing arrangement makes it hard for learners to decide which example features are relevant to solving the underlying problem (i.e., structural features) and which are not (i.e., surface features). In a structure-emphasizing arrangement, in contrast, one principle is taught using examples with different story contexts, thereby making it clear to learners that variations in surface features are irrelevant to the principle explained in the examples.

Quilici and Mayer (1996) investigated the effect of these arrangements on novice students' ability to assign novel problems to the solution principles that had been taught earlier. Given that surface features are generally more salient to inexperienced or novice learners, it was assumed that learners would focus more on surface features in situations in which surface and structural features were confounded (surface-emphasizing arrangement). In contrast, in situations in which different surface stories are used for the same structural feature (structure-

emphasizing arrangement), the authors assumed that learners would focus more on structural features. Results showed that learners who received a structure-emphasizing arrangement of examples did indeed categorize more problems according to their underlying structural features, whereas learners provided with a surface-emphasizing arrangement of examples categorized more problems according to their surface features (Quilici & Mayer, 1996). Thus, a structure-emphasizing arrangement seemed to focus the attention of learners on structural features and a surface-emphasizing arrangement on surface features.

Importantly, the context in the studies of Quilici and Mayer (1996) was completely irrelevant for learning, that is, it was only for illustrative purposes and could easily have been exchanged. Depending on the learning objectives, however, the context can also be vital for learning and thus a surface-emphasizing arrangement might be beneficial. When students are supposed to learn a scientific reasoning strategy (i.e., a principle) with examples taken from several school subjects (e.g., biology and physics), the context, that is, biology and physics, is relevant for learning. Teaching one scientific reasoning strategy using only examples from the same context might direct learners' attention to the context, that is, biology or physics. Such a surface-emphasizing or *blocked arrangement* could thus foster learners' domain knowledge in biology or physics. Teaching one scientific reasoning strategy with examples from different contexts might direct learners' attention to the underlying scientific reasoning strategy. Such a structure-emphasizing or *mixed arrangement* could thus foster scientific reasoning skills.

Taken together, the design of worked examples influences different types of learning outcomes. The instructional approach of an example addresses the question of how to introduce the abstract principle of an example. Whereas a deductive approach of presenting the abstract principle followed by examples might foster domain knowledge, an inductive approach presenting only the examples from which the principle should be inferred might foster the acquisition of skills. In addition, the arrangement of examples addresses the question of how to arrange multiple examples. Whereas a blocked arrangement with examples from the same context might foster domain knowledge, a mixed arrangement with examples from different contexts might foster scientific reasoning skills. The design of examples will be addressed in Study 2 of the present thesis.

1.3 Overview of studies and research questions

The present thesis investigates how to foster scientific reasoning through inquiry and example-based learning. More specifically, it aims at fostering students' scientific reasoning skills with video modeling examples and inquiry tasks with virtual experiments. As was shown in Section 1.1, scientific reasoning can be considered to be an intentional knowledge-seeking process, which comprises the skills implicated in generating hypotheses, designing and conducting experiments, and evaluating evidence. Since these skills are necessary for everyone in everyday life but do not develop routinely, they need to be fostered in schools. As presented in Section 1.2, there are two promising approaches to fostering scientific reasoning skills: inquiry learning, which can be considered as learning by doing or problem solving and example-based learning, which is a form of direct instruction (learning by being told). Whereas in inquiry learning students conduct experiments, make observations or collect information in order to solve a problem, in example-based learning learners study example problems with a worked-out solution rather than solving the problems themselves. Research has shown that pure inquiry learning is less effective than direct instruction such as example-based learning. However, it is yet unclear whether there might be advantages of combining inquiry learning and example-based learning. Hitherto, there was only one study using video modeling examples in an inquiry learning environment to foster scientific reasoning skills (Mulder et al., 2014). Results showed that video modeling examples enhanced scientific reasoning skills, whereas domain knowledge of learners remained quite modest. The authors, therefore, suggested improving the delivery and the design of the video modeling examples. Thus, these aspects were addressed in the two studies of the present thesis.

To foster the acquisition of scientific reasoning, I have developed a digital training program combining *video modeling examples* and *inquiry tasks with virtual experiments* to foster scientific reasoning skills. The training program was used in both studies to teach scientific reasoning strategies (and in Study 2: domain knowledge) in the context of natural sciences to students. To create the video modeling examples and the training inquiry tasks I used several virtual experiments (Gizmos, 2016) with the topic of energy in the two domains physics and biology. When the experiments were presented as video modeling examples, students were asked to watch a short video in which two models performed an inquiry task using the virtual experiments. When the experiments were presented as inquiry tasks, students were asked to conduct the same or a similar experiment that the models in the video modeling examples worked on. Thus, the same virtual experiments were used to create the video

modeling examples and the inquiry tasks. Both studies used an experimental pretest – training – posttest design. In the following, the specific research questions along with the aims and features of the two empirical studies within this thesis are elaborated:

I. How does the sequence of video modeling examples and inquiry tasks influence students' acquisition of scientific reasoning skills?

The first study addressed the delivery of examples, or more specifically, if and how examples and inquiry tasks should be combined to foster scientific reasoning skills. As discussed above in Section 1.2.2.3, it is yet unclear whether there are advantages of combining examples and inquiry tasks. In addition, if examples and inquiry tasks should be combined the question arises of how to sequence the two learning activities. On the one hand, there is research speaking in favor of presenting examples first since this might reduce cognitive load in learners and help them to create a problem-solving schema. Solving an inquiry task subsequently might help to stabilize and use the schema. On the other hand, there is research speaking in favor of solving inquiry tasks first before studying examples. Solving an inquiry task might help learners to recognize knowledge gaps and enhance their monitoring accuracy. Thus, they might be better prepared and more motivated to study a subsequent example more closely. Thus, Study 1 investigated the effects of four conditions (example-example, example-inquiry task, inquiry task-example and inquiry task-inquiry task) on students' acquisition of scientific reasoning skills. During school lessons, 107 seventh graders learned how to apply the control-of-variables strategy with a digital training program including virtual physics experiments. The delivery of examples varied according to condition. Effects on scientific reasoning skills were assessed with a multiple-choice scientific reasoning test and through analyzing students' experimenting behavior (number of controlled and confounded experiments). In addition, effects on cognitive load, judgments of learning, and monitoring accuracy were investigated.

II. How does the design (arrangement and instructional approach) of video modeling examples influence the acquisition of scientific reasoning skills and domain knowledge?

The second study investigated how the design of video modeling examples can be optimized to foster scientific reasoning skills and domain knowledge simultaneously. Targeted design aspects were the instructional approach and the arrangement of examples with regard to their structural and context features. Video modeling examples were designed using either a deductive (principle followed by examples) or an inductive instructional approach (only

examples from which students had to infer the principle). Moreover, video modeling examples were presented either in a blocked (examples with the same context) or a mixed arrangement (examples with different contexts). Eighth graders (N = 124) were randomly assigned to the four groups of this 2x2 design. Again, scientific reasoning skills were assessed with a multiple-choice scientific reasoning test and through analyzing students' experimenting behavior (number of controlled, confounded, and identical experiments). Domain knowledge was assessed with a multiple-choice test.

2 Study 1

Study 1[1] investigated how the sequence of video modeling examples and inquiry tasks influences students' acquisition of scientific reasoning skills. More precisely, it was investigated whether video modeling examples in a simulation-based inquiry learning environment should be provided before, after, or instead of an inquiry task. The learning environment consisted of two virtual physics experiments on the topic of energy. When the experiments were presented as video modeling examples, learners watched a video showing how two models solved an inquiry task. When the experiments were presented as inquiry tasks, learners had to solve the same inquiry task as the models on their own. Learners received either an example or inquiry task in a first training phase followed by an example or inquiry task in a second training phase. I compared the four resulting instructional conditions of this 2x2 design with regard to cognitive load, learning outcomes, judgments of learning (JoLs), and monitoring accuracy. In line with research on the worked example effect and on sequencing instruction and inquiry (Barzilai & Blau, 2014; Lazonder et al., 2010; Leppink et al., 2014; van Gog et al., 2011), I hypothesized that watching a video modeling example first would be more effective (better learning outcomes; Hypothesis 1a) and efficient (lower cognitive load; Hypothesis 1b) than solving an inquiry task first. Moreover, I investigated whether the instructional conditions would affect learners' JoLs. In line with research on worked examples and monitoring accuracy as well as on preparation for future learning/productive failure (Arena & Schwartz, 2014; Baars et al., 2016; Baars, van Gog, et al., 2014; Loibl & Rummel, 2014), I expected that the example-first groups would be overconfident in their JoLs in the first training phase (Hypothesis 2a), whereas the task-first groups would have more accurate JoLs (Hypothesis 2b). Moreover, I hypothesized that solving an inquiry task after having watched a video modeling example would lead to more accurate monitoring after the second training phase (Hypothesis 2c), whereas studying a second video modeling example would lead to overconfidence (Hypothesis 2d).

[1] The results of this study have been submitted as Kant, J., Scheiter, K. & Oschatz, K. (under revision). How to Sequence Video Modeling Examples and Inquiry Tasks to Foster Scientific Reasoning. *Learning and Instruction.*

2.1 Method

2.1.1 Participants and design

Participants were 107 German high school students from Grade 7 from two schools in Southern Germany (61 female, age M = 12.46 years, SD = 0.56). Participants were enrolled in their first course of physics. Because data collection took place in the beginning of the school term, students were assumed to be novices concerning the topic of energy in physics. Participation in the study was voluntarily and written informed consent from parents and children was obtained. All participants engaged in two training phases in each of which they learned the CVS with virtual experiments in the form of either video modeling examples or inquiry tasks (see Figure 1).

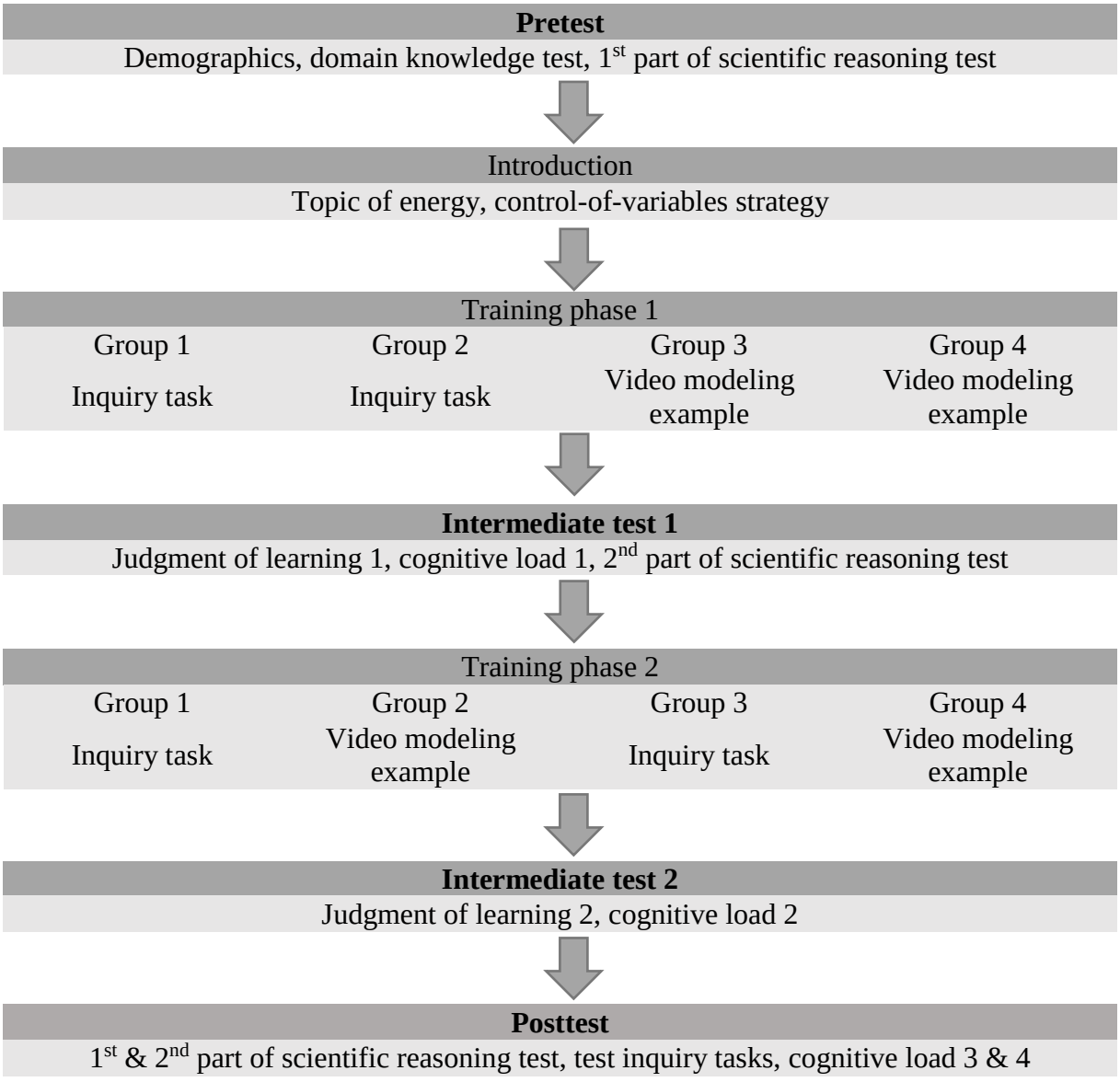

Figure 1. Procedure with the four assessment time points (in bold) and the assessed variables.

Participants were randomly assigned to one of four conditions: inquiry tasks only (task-task, *n* = 27), inquiry task followed by video modeling example (task-example, *n* = 26), video modeling example followed by inquiry task (example-task, *n* = 27), and video modeling examples only (example-example, *n* = 27).

2.1.2 Materials

2.1.2.1 Learning content. At the beginning, all participants received a short, written introduction into the topic of energy as well as an abstract description of the CVS. Afterwards, participants learned the strategy with concrete virtual experiments presented either as video modeling examples or as training inquiry tasks in two subsequent training phases. To create the video modeling examples and the training inquiry tasks I used two virtual experiments: Heat Absorption and Energy Conversion in a System (Gizmos, 2016). In the simulation called Heat Absorption, students could shine a flashlight on a variety of materials, and measure how quickly each material heats up. Students could vary the light angle, light color, type of material, and color of material and investigate their influence on the heating of the material (see Figure 2).

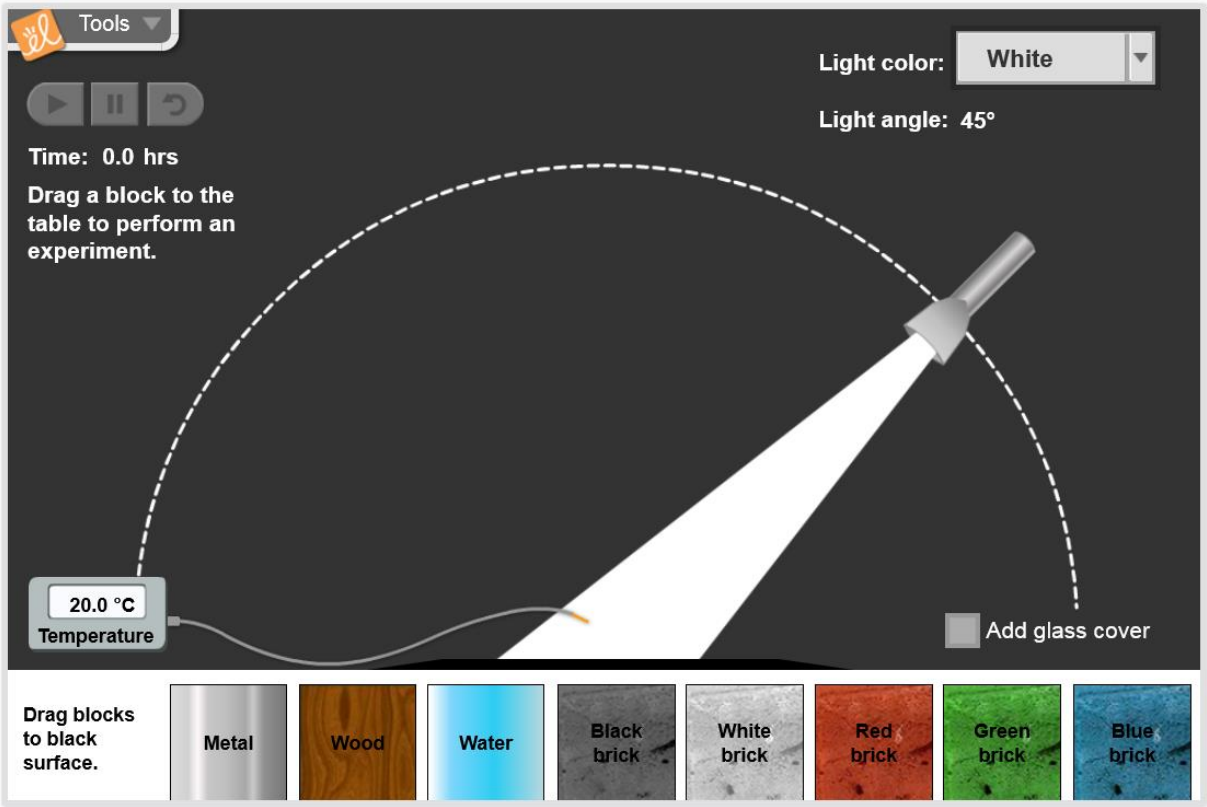

Figure 2. Screenshot of virtual experiment on heat absorption. Copyright (2016) by ExploreLearning. Reprinted with permission.

In Energy Conversion in a System, a falling cylinder was attached to a rotating propeller that stirred and heated the water in a beaker. Students could adjust the mass and height of the cylinder as well as the quantity and initial temperature of water to determine the temperature of the water as energy was converted from one form to another.

2.1.2.2 Video modeling examples. When the experiments were presented as video modeling examples, students were asked to watch a short video in which two models performed an inquiry task using the virtual experiments. The videos included a screen capture recorded with Camtasia Studio 8.5 of what the models saw on the screen while they interacted with the virtual experiments as well as verbal comments of the models describing their reasoning. Only the results of the models' actions were visible in the video but not the models themselves. In both video modeling examples, the models received a research question (e.g., in Heat Absorption: 'How does the light angle influence the heating of a material?'). The models first investigated the setup of the experiment and checked which variables they could vary. Afterwards, they tried to investigate the research question using the CVS. I used coping models, that is, the models' initial performance included errors that they identified and then corrected (van Gog & Rummel, 2010). The video modeling examples stopped several times and students were asked about critical aspects of the examples. For example, they were prompted to explain to themselves the connection between the abstract strategy description and the modeling examples (e.g., 'Please describe what Hanna and Tom did to obtain results that can unambiguously be interpreted.').

2.1.2.3 Training inquiry tasks. When the experiments were presented as inquiry tasks, students were asked to conduct an experiment with the same research questions that the models in the video modeling examples worked on. To ensure comparability of conditions, students with training inquiry tasks also received the first part of the video modeling example, where the models investigated and explained the setup of the experiment. Subsequently, students were asked to use the CVS to investigate the research question on their own. They were guided during their inquiry with analogous prompts as students with video modeling examples (e.g., 'Please describe what you did to obtain results that can unambiguously be interpreted.').

2.1.3 Measures

There were four assessment time points at which various measures were assessed: one prior to learning (pretest), one between the first and the second training phase to assess effects

of the first training phase (intermediate test 1), one after the second training phase (intermediate test 2), and a final test phase after learning (posttest).

Prior to learning, I assessed participants' age, gender, grades in biology and math as well as interest and self-efficacy in physics as control variables.[2] Moreover, I assessed the domain knowledge of participants with six self-developed multiple-choice items about the topic of energy in physics. All items consisted of a question (e.g., 'What is the law of energy conservation?') with four answer options. Participants were asked to choose their answer and were given one point for each correct answer. I calculated the percentage of correct answers for each student.

During the intermediate tests and the posttest, I assessed the dependent variables scientific reasoning, cognitive load, and JoLs. In the following sections, the instruments used to assess the dependent variables are described as well as the respective assessment time points.

2.1.3.1 Scientific reasoning. Scientific reasoning was assessed with two different measures: a scientific reasoning test with multiple-choice items and two test inquiry tasks. The item pool for the scientific reasoning test was based on a test developed by Koenen (2014) which assessed the ability to apply knowledge about experimental scientific practices. Each of the 18 items I used consisted of a short informational text about two students who are in a certain situation (e.g., baking muffins from batter that contain different baking agents), followed by a question (e.g., 'Regarding the muffins that result from the different batters, which conclusion about the baking agents is correct?'). Students had to choose the correct answer out of four answer options and received one point for each correct response. Percentage of correct answers for each student was calculated. Based on the item difficulties of a previous study, I split the test in two parts with nine items each that were equally difficult. The first part was used in the pretest (*Cronbach's* α = .56), whereas the second part was used in intermediate test 1 (*Cronbach's* α = .62), and the complete test was used in the posttest (*Cronbach's* α = .71).

To have a measure for scientific reasoning that is close to real experimenting behavior, I additionally administered two test inquiry tasks in the posttest. The first test inquiry task used the same virtual experiment as the second training inquiry task. However, this time students had to investigate the influence of a different dependent variable. The second inquiry task used

[2] These variables were assessed only to ensure the comparability of conditions regarding students' entry characteristics and will not be considered any further.

a virtual experiment in the domain of biology, namely photosynthesis. Students could investigate the influence of light intensity, carbon dioxide level, temperature, and wavelength of light on the photosynthesis rate of an aquatic plant. The test inquiry task was to investigate how light intensity influences the photosynthesis rate. To analyze students' performance in the test inquiry tasks, I video-recorded the computer screens of students while they experimented. Two independent coders were trained with data of a previous study (Krippendorff's α inter-rater reliability between .96 and 1) to analyze the videos using a rubric that was based on the features of Gobert et al. (2013). The rubric contained the following categories:

a) Controlled experiments with hypothesis: Coders counted the pairs of subsequent trials in which only the values of the independent variable of the hypothesis were manipulated from trial to trial while all other variables were kept equal.
b) Confounded experiments: Coders counted the number of pairs of subsequent trials in which the values of more than one variable had been manipulated between trials.

2.1.3.2 Cognitive load. Cognitive load was assessed with the two items perceived difficulty ('How easy or difficult was it to understand the learning content overall?') and mental effort ('How much effort did you invest in processing the learning material overall?', adapted from Kühl, Scheiter, Gerjets, & Gemballa, 2011; and Schmidt-Weigand & Scheiter, 2011). Participants answered on a 7-point Likert scale ranging from 1 = *very easy/no effort at all* to 7 = *very difficult/a lot of effort*. Cognitive load items were assessed in intermediate test 1 (cognitive load 1), in intermediate test 2 (cognitive load 2) and after each test inquiry task in the posttest (cognitive load 3 & 4). Cognitive load ratings 3 and 4 were averaged.

2.1.3.3 Judgments of learning. Participants were asked to make a JoL after each video modeling example or training inquiry task by responding on a scale from 0% to 100% to the following question: 'How confident are you that you will be able to correctly answer questions on the topic of experimenting based on the video modeling example/the training inquiry task you just saw?' JoL 1 was assessed in intermediate test 1 and JoL 2 in intermediate test 2.

2.1.3.4 Monitoring accuracy. To examine absolute monitoring accuracy of participants I calculated bias scores (Baars, van Gog, et al., 2014). Monitoring accuracy for intermediate test 1 was calculated by subtracting performance in the second half of the scientific reasoning test from JoL 1. Values close to zero indicate accurate judgments of learning, whereas negative values indicate underconfidence, and positive values overconfidence. Monitoring accuracy for

intermediate test 2 was calculated by subtracting performance in the complete scientific reasoning test from JoL 2.

2.1.4 Procedure

The experiment was conducted in a computer room of the participants' schools. Students within each class were randomly assigned to the experimental conditions. All participants engaged in two sessions that were scheduled across two consecutive days. On the first day, the pretest, introduction, training phase 1, intermediate test 1, training phase 2, and intermediate test 2 took place. On the second day, participants completed the posttest. Due to scheduling problems, one class completed the posttest with a delay of two days instead of one. During the experiment, participants worked individually on computers. Each participant had a headset for listening to the comments of the models in the video modeling examples. At the beginning of the first day, I assessed demographic data, domain knowledge, and scientific reasoning (pretest). This phase took roughly 15 minutes. Then, students received a written introduction to the topic of energy and experiments including the CVS. In training phase 1, students watched a first modeling example or conducted a first training inquiry task, depending on condition. Subsequently, in intermediate test 1, students gave a JoL, rated their cognitive load, and worked on the second part of the scientific reasoning test. In the subsequent training phase 2, students watched a second video modeling example or conducted a second training inquiry task. Afterwards, in intermediate test 2, participants again gave a JoL and rated their cognitive load. The complete procedure from introduction to intermediate test 2 took on average 51 minutes. On the second day, all participants completed the posttest that contained the complete scientific reasoning test as well as the two test inquiry tasks with cognitive load measures.

2.2 Results

As a check of randomization, the domain knowledge test and scientific reasoning pretest were compared with a 2x2 MANOVA with first format (video modeling example vs. inquiry task) and second format (video modeling example vs. inquiry task) as factors. Results showed – as expected – no significant differences between conditions, effect of first format: $F(2, 102) = 1.06$, $p = .35$, $\eta_p^2 = .02$, effect of second format: $F(2, 102) = 1.53$, $p = .22$, $\eta_p^2 = .03$, and interaction effect: $F(2, 102) = 1.00$, $p = .37$, $\eta_p^2 = .01$. In Table 1 the pretest data as well as the data of the intermediate tests and the posttest is presented per condition.

2.2.1 Intermediate test 1

To examine the effects of the first format (video modeling example vs. inquiry task), I aggregated groups 1 and 2 as well as groups 3 and 4, since they had both solved the same inquiry task or watched the same video modeling example, respectively. Thus, in intermediate test 1, there were effectively only two intervention groups.[3] One person with missing values was excluded from the analyses. For group comparisons, ANOVAs and MANOVAs with format (video modeling example vs. inquiry task) as the factor were conducted.

2.2.1.1 Scientific reasoning test. To test whether watching a video modeling example first instead of solving an inquiry task first would lead to better scientific reasoning, an ANOVA was run for performance on the scientific reasoning test. There was a significant effect of format on scientific reasoning performance, $F(1, 104) = 8.90$, $MSE = 424.98$, $p = .004$, $\eta_p^2 = .08$, corresponding to a medium effect (Cohen, 1988). Learners who watched a video modeling example first ($M = 76.73$, $SD = 19.03$) showed a higher performance in the scientific reasoning test than learners who solved an inquiry task first ($M = 64.78$, $SD = 22.09$), which corresponds to a worked example effect.

2.2.1.2 Cognitive load. I hypothesized that watching a video modeling example first would reduce cognitive load, compared to solving and inquiry task first. To test this hypothesis a MANOVA was run for subjective difficulty and mental effort. There was a significant main effect on cognitive load, $F(2, 103) = 4.27$, $p = .02$, $\eta_p^2 = .08$, corresponding to a medium effect.

[3] I compared the results with a 2x2 ANOVA with the factors first format (video modeling example vs. inquiry task) and second format (video modeling example vs. inquiry task). As expected, for all dependent variables neither second format nor the interaction of first and second format became significant since in intermediate test 1 the second format was not yet present.

Table 1

Means and standard deviations (in parentheses) as a function of experimental condition.

Time	Variables	Task-task	Task-example	Example-task	Example-example
Pretest	Domain knowledge (%)	36.42 (19.08)	42.31 (15.80)	38.89 (15.33)	43.83 (14.73)
	Scientific reasoning (%)	62.55 (17.20)	66.24 (22.11)	72.84 (19.08)	65.84 (19.71)
Intermediate test 1	Subjective difficulty (1-7)	3.30 (1.07)	3.31 (1.29)	2.81 (1.11)	3.54 (1.50)
	Mental effort (1-7)	5.33 (1.18)	5.27 (1.25)	4.37 (1.36)	4.85 (1.08)
	Judgment of learning (%)	56.15 (24.70)	60.38 (25.49)	64.19 (23.33)	63.92 (24.14)
	Scientific reasoning (%)	65.02 (21.06)	64.53 (23.52)	78.19 (18.61)	75.21 (19.70)
	Monitoring accuracy	-8.87 (27.71)	-4.15 (36.72)	-14.00 (27.00)	-11.29 (32.59)
Intermediate test 2	Subjective difficulty (1-7)	3.81 (1.91)	3.96 (1.52)	3.05 (1.31)	3.72 (1.28)
	Mental effort (1-7)	5.05 (1.83)	5.26 (1.25)	4.42 (1.58)	4.77 (1.45)
	Judgment of learning (%)	49.00 (27.50)	56.00 (25.02)	68.42 (21.74)	64.82 (21.20)
	Monitoring accuracy	-24.01 (21.68)	-12.84 (30.84)	-11.99 (19.19)	-10.94 (24.80)
Posttest	Scientific reasoning (%)	72.43 (16.49)	71.79 (20.30)	79.01 (14.64)	77.98 (12.56)
	Test inquiry tasks				
	No. of controlled experiments	5.56 (4.89)	8.77 (5.85)	10.19 (7.87)	11.88 (5.17)
	No. of confounded experiments	0.85 (1.26)	1.85 (2.03)	0.38 (0.90)	0.92 (1.55)
	Subjective difficulty (1-7)	2.83 (1.39)	2.62 (1.94)	2.21 (0.93)	2.67 (0.99)
	Mental effort (1-7)	4.78 (1.81)	4.63 (1.58)	4.27 (1.91)	4.12 (1.73)

Follow-up ANOVAs showed that learners who watched a video modeling example (M = 4.60, SD = 1.65) reported less mental effort during the first training phase than learners who solved an inquiry task (M = 5.30, SD = 1.20), $F(1, 104) = 8.62$, MSE = 1.50, p = .004, η_p^2 = .08. There were no significant differences in subjective difficulty between the groups (Fs < 1).

2.2.1.1 Judgments of learning. An ANOVA revealed no significant effect on JoL 1, $F(1, 104) = 1.53$, MSE = 587.17, p = .22, η_p^2 = .02.

2.2.1.2 Monitoring accuracy. I hypothesized that watching a video modeling example first would lead to overconfidence, whereas solving an inquiry task first would lead to more accurate JoLs. To test this hypothesis, I first conducted one-sample t-tests against zero for the monitoring accuracy in both groups. Mean monitoring accuracy of participants who watched a video modeling example first (M = -12.67, SD = 29.61) was significantly different from zero, $t(53) = -3.12$, p = .003. However, in contrast to the hypothesis, participants who watched a video modeling example significantly underestimated rather than overestimated themselves. Mean monitoring accuracy of participants who solved an inquiry task (M = -6.55, SD = 32.21) was not significantly different from zero, $t(53) = -1.48$, p = .15. Thus, in line with my hypothesis, participants who solved an inquiry task first gave accurate JoLs. Second, an ANOVA was run to test whether the groups differed in their monitoring accuracy, revealing no significant effect, $F(1, 104) = 1.04$, MSE = 957.38, p = .31, η_p^2 = .01.

2.2.2 Intermediate test 2

To examine effects of the four instructional conditions on the dependent variables in intermediate test 2, I performed 2x2 (M)ANOVAs with first format (video modeling examples vs. inquiry task) and second format (video modeling example vs. inquiry task) as factors (see Leppink et al., 2014). This enabled me to test (1) the effect of first format (i.e., main effect of first format), (2) the effect of second format (i.e. main effect of second format), and (3) any extra effects of one specific condition (i.e., interaction effect of first format and second format). Because of time constraints in schools, 22 participants did not complete intermediate test 2. After determining that these missing values were independent of group membership (χ^2 (3) = 2.77, p = .43), I excluded these participants from the analyses for intermediate test 2.

2.2.2.1 Cognitive load. Because I expected participants who watched a video modeling example first to have built a problem-solving schema after the first training phase, I hypothesized that these participants would also report lower cognitive load in the second training phase than participants who solved an inquiry task first. I tested this hypothesis with a

2x2 MANOVA for subjective difficulty and mental effort. There was neither a significant main effect of first format on cognitive load, $F(2, 80) = 1.77$, $p = .18$, $\eta_p^2 = .04$, nor a main effect of second format, nor an interaction effect (both Fs < 1). Taken together, subjective difficulty and mental effort in intermediate test 2 did not differ between the four groups.

2.2.2.2 Judgments of learning. To explore whether there would be differences in JoLs in intermediate test 2 between the four groups, a 2x2 ANOVA was run for JoL 2. There was a significant main effect of first format, $F(1, 81) = 7.29$, $MSE = 578.22$, $p = .01$, $\eta_p^2 = .08$, corresponding to a medium effect. Learners who watched a video modeling example first (M = 66.49, SD = 21.26) reported higher JoLs than learners who solved an inquiry task first (M = 52.66, SD = 26.16). There was neither a significant main effect of second format ($F < 1$), nor an interaction effect, $F(1, 81) = 1.03$, $MSE = 578.22$, $p = .31$, $\eta_p^2 = .01$. Therefore, I can conclude that participants in the example-task group and in the example-example group were more confident about their future scientific reasoning performance after the second training phase than participants in the task-task group and the task-example group.

2.2.2.3 Monitoring accuracy. I hypothesized that participants in the example-task group would be accurate in their JoLs, whereas the example-example group would be overconfident. Moreover, I explored whether there would be any effect on monitoring accuracy of participants in the task-task and the task-example group. For this purpose, I again first conducted one-sample t-tests against zero for all groups. Participants in the example-task group (M = -11.99, SD = 19.19) were not accurate but underconfident, $t(18) = -2.72$, $p = .01$. The same was true for participants in the example-example group (M = -10.93, SD = 24.80, $t(21) = -2.07$, $p = .05$) and in the task-task group, M = -24.02, SD = 21.68, $t(20) = -5.08$, $p < .001$. Finally, for participants in the task-example group (M = -12.84, SD = 30.84) there was a tendency to be underconfident, $t(22) = -2.00$, $p = .06$. Taken together, all groups underestimated their ability to correctly answer questions on the topic of experimenting.

Second, I investigated whether there were differences in monitoring accuracy of participants by means of a 2x2 ANOVA. There were no main effects of either first or second format on monitoring accuracy ($F(1, 81) = 1.67$, $MSE = 615.67$, $p = .20$, $\eta_p^2 = .02$ and $F(1, 81) = 1.28$, $MSE = 615.67$, $p = .26$, $\eta_p^2 = .02$, respectively) nor an interaction effect ($F < 1$).

2.2.3 Posttest

Two students did not complete the test inquiry tasks and thus had missing values for inquiry tasks and cognitive load. They were excluded from the respective analyses.

2.2.3.1 Scientific reasoning test. Because I expected participants who watched a video modeling example first to have built a problem-solving schema after the first training phase, I hypothesized that these participants would also have a higher scientific reasoning performance on the posttest than participants who solved an inquiry task first. To test this hypothesis, I conducted a 2x2 ANOVA with first format (video modeling example vs. inquiry task) and second format (video modeling example vs. inquiry task) for scientific reasoning performance on the posttest. There was a significant main effect of first format, $F(1, 103) = 4.15$, $MSE = 262.63$, $p = .04$, $\eta_p^2 = .04$, corresponding to a small effect. Analogously to intermediate test 1, learners who watched a video modeling example first ($M = 78.50$, $SD = 13.52$) showed a higher performance in the scientific reasoning posttest than learners who solved an inquiry task first ($M = 72.12$, $SD = 18.28$). There was neither a significant main effect of second format nor an interaction effect (Fs < 1). In sum, students in the example-task and in the example-example group performed better in the final scientific reasoning posttest than students in the task-task and in the task-example group.

2.2.3.2 Test inquiry tasks. Analogously to the scientific reasoning test performance, I hypothesized that participants who watched a video modeling example first would perform better in solving the test inquiry tasks than participants who solved an inquiry task first. This should be reflected in a higher number of controlled and a lower number of confounded experiments. To test this hypothesis, I conducted a 2x2 MANOVA for the number of controlled and confounded experiments. There was a significant main effect of first format on performance in the test inquiry tasks, $F(2, 100) = 6.68$, $p = .002$, $\eta_p^2 = .12$, corresponding to a medium to large effect. Follow-up ANOVAs showed that participants who watched a video modeling example first ($M = 11.04$, $SD = 6.65$) performed more controlled experiments than participants who solved an inquiry task first ($M = 7.16$, $SD = 5.57$), $F(1, 101) = 10.78$, $MSE = 36.57$, $p = .001$, $\eta_p^2 = .10$ (medium effect). Moreover, participants who watched a video modeling example first also conducted fewer confounded experiments ($M = 0.65$, $SD = 1.28$) than participants who solved an inquiry task first ($M = 1.34$, $SD = 1.74$), $F(1, 101) = 5.70$, $MSE = 2.23$, $p = .02$, $\eta_p^2 = .05$ (small effect). Additionally, there was a significant main effect of second format $F(2, 100) = 7.28$, $p = .001$, $\eta_p^2 = .13$, corresponding to a medium to large effect. Follow-up ANOVAs showed that participants who watched a video modeling example second ($M = 10.33$, $SD = 5.69$) performed more controlled experiments than participants who solved an inquiry task second ($M = 7.83$, $SD = 6.87$), $F(1, 101) = 4.32$, $MSE = 36.57$, $p = .04$, $\eta_p^2 = .04$ (small effect). However, participants who watched a video modeling example second also conducted more confounded experiments ($M = 1.38$, $SD = 1.85$) than participants who solved

an inquiry task second (M = 0.62, SD = 1.11), $F(1, 101) = 6.93$, MSE = 2.23, p = .01, η_p^2 = .06 (small effect). There was no significant interaction, $F < 1$.

2.2.3.3 Cognitive load. Analogously to intermediate test 1, I hypothesized that participants who watched a video modeling example first instead of solving an inquiry task first would experience lower cognitive load in posttest. A MANOVA for subjective difficulty and mental effort revealed neither a main effect of first format, ($F(2, 100) = 1.55$, p = .22, η_p^2 = .03), nor a main effect of second format, ($F < 1$), nor an interaction, $F(2, 100) = 1.22$, p = .30, η_p^2 = .02. In sum, subjective difficulty and mental effort during posttest did not differ between groups.

2.3 Discussion

The present study investigated whether video modeling examples in simulation-based inquiry learning should be provided before, after, or instead of inquiry tasks. Learners were either provided with a video modeling example or an inquiry task in a first training phase and a video modeling example or an inquiry task in a second training phase. I analyzed the effects on learning outcomes, cognitive load, judgments of learning (JoLs), and monitoring accuracy. Results indicated an advantage for providing video modeling examples before or instead of but not after an inquiry task. Participants who watched a video modeling example first reported less mental effort in the first training phase and showed better learning outcomes in the first intermediate test as well as in the posttest. Moreover, they were more confident about their future test performance after the second training phase. However, all groups underestimated their performance.

2.3.1 Hypothesis 1: Worked example effect

The results for learning outcomes and cognitive load are in line with hypothesis 1 and indicate a worked example effect. For learners with little prior knowledge it was beneficial to watch a video modeling example first. Studying this example reduced learners' cognitive load and seems to have helped them to concentrate on the steps that are necessary to solve an inquiry task. Thus, they possibly were able to build a problem-solving schema which not only helped them to apply knowledge about experimental scientific practices in the scientific reasoning test but also to solve inquiry tasks on their own. Interestingly, this advantage manifested itself already after the first training phase, where students who had watched a video modeling example reported lower mental effort and showed higher performance than participants who had solved an inquiry task. Thus, one example already helped to build a cognitive schema. This finding is in line with research showing that learners do not always need to study multiple examples to acquire cognitive schemata (Scheiter, Gerjets, & Schuh, 2004).

Regarding cognitive load, I found significant differences between groups only for invested mental effort but not for perceived difficulty of the learning contents. Therefore, the format of the simulated experiments (video modeling examples vs. inquiry tasks) did not make the learning content, that is, the control-of-variables strategy less or more difficult to understand. Rather it influenced how much mental effort the learners put into understanding it.

However, I did not find the expected differences in cognitive load for the second training phase or the posttest. This seems to implicate that learners had to invest less mental effort to

build a cognitive schema for later problem-solving but that they still had to invest mental effort to further stabilize and use this problem-solving schema. However, in the second intermediate test there was at least a descriptive difference in mental effort between the groups with less invested mental effort in the example-first groups. Therefore, the lack of statistical difference could also be due to a lack of power.

Nevertheless, studying an example first also helped learners in the posttest to answer scientific reasoning test items and to solve inquiry tasks on their own. Here again, participants who watched a video modeling example first showed a higher performance in the scientific reasoning test than participants who solved an inquiry task first. Moreover, the example-first groups outperformed the task-first groups in solving inquiry tasks in that they conducted more controlled and fewer confounded experiments.

There were no effects of second format on cognitive load but unexpectedly on performance in the test inquiry tasks. Learners who watched a video modeling example second conducted more controlled but also more confounded experiments than participants who solved an inquiry task second. Importantly, this effect was independent of the first format. It seems that watching a video modeling example in the second phase led learners to conduct more experiments overall (controlled and confounded) during the posttest one day later. Understanding the reason for this finding will require further investigation in future studies.

2.3.2 Hypothesis 2: Monitoring accuracy

In contrast to other studies (Baars et al., 2016; Baars, van Gog, et al., 2014), I did not find that studying an example led to inaccurate monitoring and specifically overconfidence. On the contrary, participants who had watched a video modeling example in the first training phase significantly underestimated their scientific reasoning performance. This was not in line with hypothesis 2a. One explanation for this result could be that scientific reasoning was very difficult for students so that they felt overwhelmed and not very confident about their future test performance. However, learners perceived the learning material overall as rather easy. Additionally, learners who solved an inquiry task first were rather accurate in their JoLs, which confirms hypothesis 2b. This speaks against too difficult learning material.

In the second intermediate test, all of the groups underestimated themselves even if the mean of the task-example group differed only marginally from zero. This was not in line with my hypotheses that the example-task group would accurately monitor itself (hypothesis 2c), whereas the example-example group would overestimate itself (hypothesis 2d). Therefore, the

question remains why all of the groups underestimated themselves. There are at least two (not mutually exclusive) possible explanations: (1) students underestimated themselves because of sophisticated epistemological beliefs regarding scientific reasoning, or (2) because of an underconfidence-with-practice effect (Koriat, Sheffer, & Ma'ayan, 2002). Regarding the first explanation, the topic of scientific reasoning is closely related to understanding the nature of science, that is, students' epistemological beliefs about knowledge and knowing in science (J. Mayer, 2007). One important dimension of epistemological beliefs is certainty of knowledge and knowing in science (Conley, Pintrich, Vekiri, & Harrison, 2004). Whereas less sophisticated stances would reflect the belief in a right answer, beliefs that are more sophisticated would acknowledge that there might be more than one answer to complex problems. If the students held sophisticated beliefs about the nature of science, this could have resulted in an underestimation of their own performance simply because they thought it is hard to find answers to questions in science in general. Further research should assess the epistemological beliefs of students to test this possible post-hoc explanation.

Regarding the second explanation, research on learning word pairs has shown that repeated study-test cycles lead to underconfidence in learners from the second cycle on (e.g., Koriat et al., 2002). One explanation for this effect is that learners might rely on a memory-for-past-test heuristic (Finn & Metcalfe, 2007). When learners remember that they were not able to retrieve a given item on a past test, they are not confident that they will be able to retrieve it in a future test. However, they disregard the fact that there is another training phase before the next test which might enhance their future test performance (Finn & Metcalfe, 2007). Learners in this study might also have based their JoLs on this memory-for-past-retrieval heuristic. Participants who watched a video modeling example first had a higher scientific reasoning performance in the first intermediate test than participants who solved an inquiry task first. The experience with the first intermediate test might have influenced their JoLs in the second intermediate test which may have resulted in higher JoLs in the example-first groups compared to the task-first groups. However, all participants neglected the potential knowledge gains from the second training phase which resulted in underestimation of their performance in the second intermediate test.

2.3.3 Conclusion

In conclusion, video modeling examples were found to improve scientific reasoning skills in students when they were provided before or instead of solving inquiry tasks. This seems to implicate that learners with low prior knowledge need guidance especially at the beginning

of a new learning episode. Afterwards, solving an inquiry task or watching a second video modeling example seem to be equally effective and efficient to foster scientific reasoning skills.

Since there was a relatively clear advantage of example-based learning, the second study focused on the design of video modeling examples. In schools, there are seldom single learning goals for a lesson or a teaching unit such as acquiring the CVS but rather multiple learning goals such as fostering the acquisition of scientific reasoning skills and domain knowledge simultaneously. Consequently, the second study of this thesis investigated how to design video modeling examples to foster the acquisition of scientific reasoning skills and domain knowledge simultaneously.

3 Study 2

Study 2[4] investigated whether an optimized design of modeling examples could foster scientific reasoning and domain knowledge simultaneously in a simulation-based inquiry-learning environment. The learning environment consisted of four virtual experiments, all dealing with the topic of energy in the domains of either biology or physics. Additionally, video modeling examples were part of the learning environment. In these examples, it was shown how virtual experiments should be conducted (control-of-variables strategy and hypothesis generation). I tested the effects of four different design versions of the video modeling examples. First, the examples were presented either using a deductive approach (principle followed by examples) or an inductive approach (examples only). In line with previous research (Renkl, 2015; Ross & Kilbane, 1997), I expected the inductive approach to foster the acquisition of scientific reasoning skills (Hypothesis 1a), whereas the deductive approach would foster declarative domain knowledge (Hypothesis 1b). Second, I varied the arrangement of the video modeling examples. Participants learned one scientific reasoning strategy either with two examples from different contexts (biology and physics; mixed arrangement) or with two examples from similar contexts (either biology or physics; blocked arrangement). In line with research of Quilici and Mayer (1996), I hypothesized that the mixed arrangement would foster the acquisition of scientific reasoning skills (Hypothesis 2a). The blocked arrangement, in contrast, was expected to foster domain knowledge (Hypothesis 2b). In addition, I explored possible interactions between the two factors to figure out whether one factor could compensate for the other.

[4] The results of this study will be submitted as Kant, J., Scheiter, K. & Oschatz, K. Fostering Scientific Reasoning with Video Modeling Examples of Simulated Experiments. *Instructional Science*.

3.1 Method

3.1.1 Participants and design

Participants were 126 German high school students from two schools in Southern Germany. I had to exclude two participants due to technical problems. Student absenteeism further reduced the sample to 118 students (55 female). The average age of the students was 13.38 years (*SD* = 0.55). Participation in the study was voluntary, and I obtained written informed consent from parents and children. I used a repeated measures 2 x 2 design with the within-subject factor time (pretest vs. posttest) and the between-subjects factors instructional approach (deductive vs. inductive) and arrangement of modeling examples (mixed vs. blocked). Participants were randomly assigned to one of the four conditions: (1) deductive instructional approach and mixed arrangement of modeling examples (deductive-mixed, *n* = 27), (2) deductive instructional approach and blocked arrangement of modeling examples (deductive-blocked, *n* = 30), (3) inductive instructional approach and mixed arrangement of modeling examples (inductive-mixed, *n* = 30), and (4) inductive instructional approach and blocked arrangement of modeling examples (inductive-blocked, *n* = 31).

3.1.2 Learning materials

The following section describes the scientific reasoning strategies that participants learned followed by a description of the virtual experiments that were used for this purpose. After that, the video modeling examples and the training inquiry tasks are introduced.

3.1.2.1 Scientific reasoning strategies. Participants learned two scientific reasoning strategies. (1) The control of variables strategy is crucial for conducting unconfounded experiments that can yield conclusive results. The strategy consists of keeping all variables but the variable of interest constant in order to be able to unambiguously determine how the variable of interest influences a dependent variable (Chen & Klahr, 1999). (2) The generation of hypotheses is another important aspect of scientific reasoning (de Jong & van Joolingen, 1998). Hypotheses should describe the relationship between two variables in a way that can be falsified with an experiment. The variables as well as the relationship should be measurable. These two scientific reasoning strategies were conveyed to students using video modeling examples.

3.1.2.2 Virtual experiments. To create the modeling examples and the training inquiry tasks, I used four virtual experiments (Gizmos, 2016) dealing with the topic of energy. Two virtual experiments stemmed from the domain of biology: *Photosynthesis Lab* and *Plants and Snails*, whereas the remaining two stemmed from the domain of physics: *Heat Absorption* and

Energy Conversion in a System. In the Photosynthesis Lab, students could study photosynthesis under a variety of conditions. On the left-hand side of the screen, students saw an aquatic plant within an aquarium (see Figure 3). Attached to the aquarium was a measurement device that assessed the amount of oxygen the plant produces as an indicator of the photosynthesis rate. With sliders, students could vary light intensity, carbon dioxide level, temperature and wavelength of light and investigate the influence of these independent variables on the photosynthesis rate. On the right-hand side of the screen, students could record data from their trials and display the results as a table, bar chart, or graph.

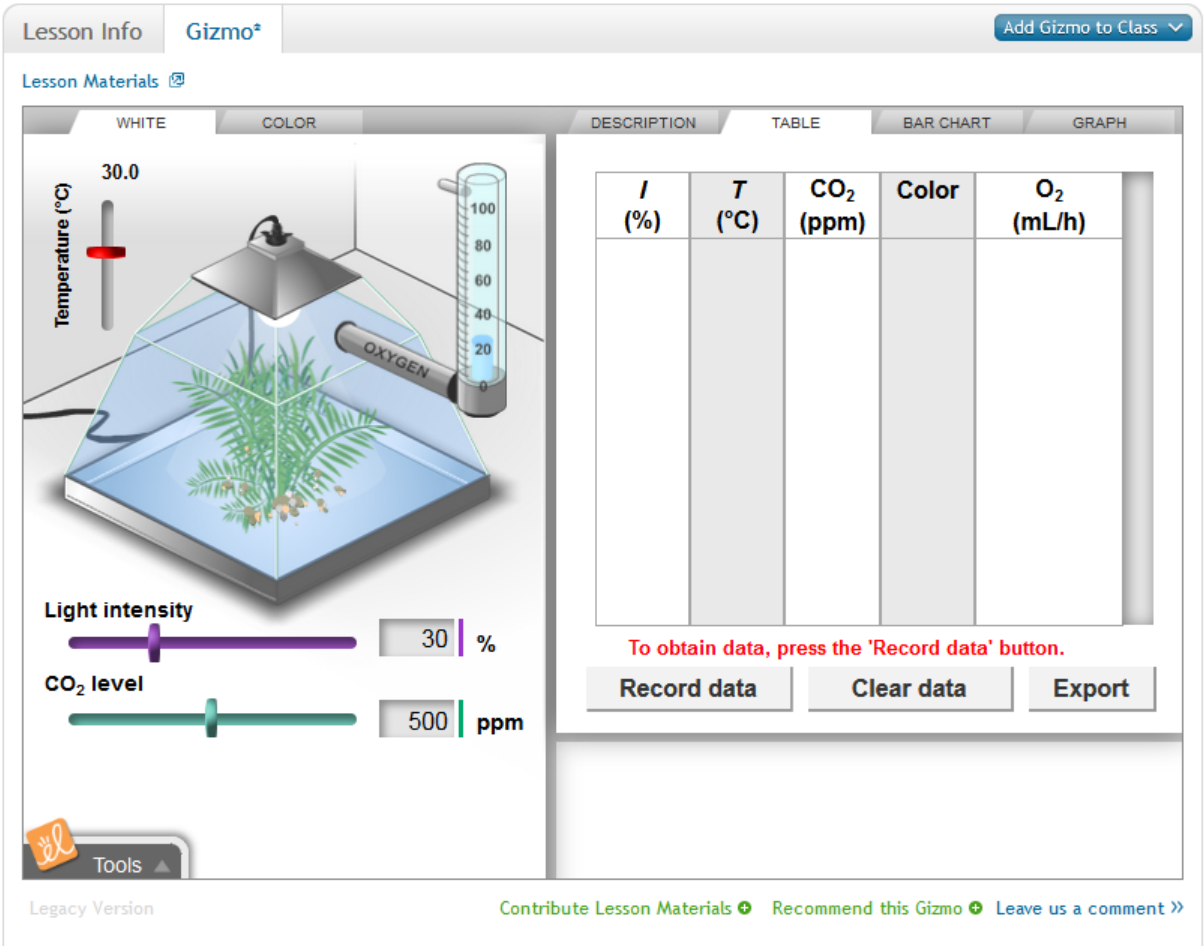

Figure 3. Screenshot of virtual experiment Photosynthesis Lab in biology. Copyright (2016) by ExploreLearning. Reprinted with permission.

The setup of the other virtual experiments was similar to the Photosynthesis Lab. In Plants and Snails, students could manipulate the number of aquatic plants and snails as well as the amount of light in test tubes and investigate the effect of these variables on the amount of oxygen and carbon dioxide in the test tubes. In Heat Absorption, students could shine a flashlight on a variety of materials and measure how quickly each material heats up as a function

of the light angle, light color, type of material, and material color. In Energy Conversion in a System, a falling cylinder was attached to a rotating propeller that stirred and heated the water in a beaker. The mass and height of the cylinder as well as the quantity and initial temperature of the water could be adjusted to determine the temperature of the water as energy is converted from one form to another.

3.1.2.3 Modeling examples. The modeling examples were short videos in which two models performed an inquiry task. The videos included a screen capture recorded with Camtasia Studio 8.5 of what the models saw on the screen while they interacted with the virtual experiments as well as their verbal comments describing their reasoning. Two amateur actors (one male, one female) served as models. They were approximately of the same age as the participants. The models were not visible in the video, only the results of their actions. The models acted according to a script. In all of the modeling examples, the models were provided with a research question (e.g., in the Photosynthesis Lab: 'How does the amount of carbon dioxide influence the oxygen production of the plant?'). The models first investigated the setup of the experiment and checked which variables could be varied. Afterwards, they tried to investigate the research question while using one of the scientific reasoning strategies, for instance, the control of variables strategy. Since I used coping models, the models made errors, recognized them and finally corrected them (van Gog & Rummel, 2010). For instance, the models would first vary carbon dioxide level and light intensity simultaneously. Then they realized that this approach yields inconclusive results. Thus, in the next trial, they varied only the values of the variable of interest, in this case carbon dioxide, and interpreted their results correctly.

The conditions differed in terms of instructional approach as well as in the arrangement of the modeling examples. The instructional approach was either deductive or inductive. In the deductive conditions, participants received an abstract description of the scientific reasoning strategy up-front in a written format followed by two modeling examples. In addition, students were prompted to explain to themselves the connection between the abstract strategy description and the modeling examples ('Please explain why the video that you just saw contained the experimental strategy described at the beginning'). In the inductive conditions, participants received no abstract description of the scientific reasoning strategy; rather, they were presented with only two modeling examples. Students were prompted to compare the two modeling examples in order to induce the strategy ('Please compare the two videos. What are the similarities and differences, especially concerning the experimental strategy?').

Furthermore, participants received the two modeling examples that were used to teach each of the two scientific reasoning strategies either in a mixed arrangement or in a blocked arrangement. Participants in the mixed conditions received one biology modeling example and one physics modeling example for each of the two scientific reasoning strategies. In other words, participants learned each of the two scientific reasoning strategies in the context of two domains (mixed arrangement). This arrangement highlighted that the strategy holds true in different domains. Participants in the blocked conditions received two modeling examples from the same domain (biology or physics) for each scientific reasoning strategy. That is, students learned each of the two scientific reasoning strategies in the context of one domain (blocked arrangement). This arrangement emphasized a relation between the strategy and the domain.

3.1.2.4 Training inquiry tasks. The two training tasks consisted of virtual experiments that were also used for the modeling examples (Photosynthesis Lab, Energy Conversion in a System). Participants were instructed to investigate the influence of another independent variable as in the modeling examples (e.g., 'Try to find out how light intensity influences oxygen production of the plant.'). Moreover, they were instructed to use the scientific reasoning strategy they had just seen in the modeling examples. Students were instructed to write down their results as well as their interpretation.

3.1.3 Measures

Scientific reasoning skills were assessed using two measures: a multiple-choice achievement test and inquiry tasks. Domain knowledge was measured using a multiple-choice achievement test and transfer tasks. In a pretest, I assessed participants' prior knowledge regarding scientific reasoning and domain knowledge with multiple-choice achievement tests. During the learning phase, I assessed scientific reasoning skills using inquiry tasks and cognitive load. In the posttest, scientific reasoning skills were assessed using the achievement test as well as inquiry tasks, and domain knowledge was assessed using the achievement test and transfer tasks.

3.1.3.1 Pretest. As control variables, I assessed participants' prior knowledge regarding scientific reasoning as well as domain knowledge.

Prior knowledge in scientific reasoning was assessed with a multiple-choice achievement test developed by Koenen (2014), assessing the ability to apply knowledge about experimental scientific practices. The items consisted of a short informational text about two students who are in a certain situation (e.g., baking muffins using batter that contained different

baking agents), followed by a question (e.g., 'Regarding the muffins that resulted from the different batters, which conclusion about the baking agents is correct?'). Participants were instructed to choose the correct answer out of four answer options. I used 18 of the original 20 items, slightly adapted them, and converted them into a digital format (*Cronbach's α* = .69). Participants' answers were scored with one point for each correct response. The number of correct responses per student was divided by the maximum number of correct responses (18) and multiplied by 100 to obtain the percentage of correct answers.

Domain knowledge was assessed with eight self-developed multiple-choice items about the topic of energy in biology and physics (*Cronbach's α* = .36). All items consisted of a question (e.g., 'What is the law of energy conservation?') with four answer options. Participants had to choose the correct answer and were given one point for each correct answer. The number of correct responses per student was divided by the maximum number of correct responses (8) and multiplied by 100 to obtain the percentage of correct answers.

3.1.3.2 Learning Phase. To assess cognitive load, I adopted two items by Cierniak, Scheiter, and Gerjets (2009): 'How easy or complex was the learning material overall?' and 'How much effort did you invest in processing the learning material overall?' Participants had to answer on a 7-point Likert scale ranging from 1 = *very easy/no effort at all* to 7 = *very complex/a lot of effort*.

To be able to analyze students' performance in the training inquiry tasks, I video-recorded the computer screens of students while they experimented with Camtasia Studio 8.5. Two independent coders were trained to analyze the videos using a rubric that was based on the features of Gobert et al. (2013). The rubric contained the following categories:

a) Controlled experiments with hypotheses: Coders counted the pairs of trials in which only the values of the variable in question (i.e., the independent variable of the hypothesis) were manipulated from trial to trial while all other variables were kept equal.
b) Identical experiments: Furthermore, coders determined the number of pairs of trials that had identical experimental setups.
c) Confounded experiments: Finally, coders counted the number of pairs of trials in which the values of more than one variable had been manipulated between trials.

The coders first scored a random selection of 20% of the data. In this sample, Krippendorff's α inter-rater reliability (Hayes & Krippendorff, 2007) for the three categories in the different inquiry tasks was between .95 and 1. The instances in the random selection where the coders differed were discussed and clarified prior to the actual coding of all training inquiry tasks by the coders.

3.1.3.3 Posttest. Scientific reasoning in the posttest was assessed with the multiple-choice achievement test that had already been used in the pretest as well as with two new test inquiry tasks. The test inquiry tasks were analogous to the training inquiry tasks but used different virtual experiments. In a biology experiment *Growing Plants*, students could investigate the growth of three common garden plants: tomatoes, beans, and turnips. They could change the amount of light each plant gets, the amount of water added each day, and the type of soil the seed is planted in. Participants were asked to investigate the following hypothesis: 'If you increase the amount of water, the height and mass of tomatoes increases.' In a physics experiment *Inclined Plane*, students could observe objects of different shapes and materials as they roll or slide down an inclined plane. The slope of the plane could be adjusted, and a variety of materials could be used for the plane. Participants were asked to investigate the following hypothesis: 'If you increase the slope of the plane, the translational energy of a ring of steel increases.' The rubric for analyzing the test inquiry tasks was the same as for the training tasks. The test tasks were analyzed by the same coders (Krippendorff's α inter-rater reliability was between .85 and 1).

Domain knowledge in the posttest was assessed with the same multiple-choice achievement test already used in the pretest as well as two additional self-developed items. I had to exclude one item in the posttest due to technical problems, resulting in a total of nine items. Moreover, I used two transfer tasks adapted from KMK (2005) to assess domain knowledge. Students were provided with a problem description (e.g., about a sealed aquarium with waterweed and two great pond snails in it) and four tasks (e.g., 'After three months, the snails didn't succeed in grazing the waterweed. Instead, the waterweeds grew steadily. A student claims that the reason for this growth is the dung of the snails. The dung serves as nourishment for the plants. Please comment on this statement.'). Students' written answers were scored using a rubric from the KMK (2005). Participants were given one point for each correct concept they mentioned in their answers. The number of correct responses per student was divided by the maximum number of correct responses (16) and multiplied by 100 to obtain the percentage of correct answers.

3.1.4 Procedure

The experiment was conducted in a computer room at the participants' schools with complete classes, with the students in each class randomly assigned to the experimental conditions. All participants engaged in four lessons scheduled across three consecutive days. During the experiment, participants worked individually at the computers. On the first day, participants had a double lesson. At the beginning of this lesson, I assessed participants' age, gender, school and class. Students also indicated whether they had attended an advanced science course. I asked participants to report their grades in biology and physics. Then I assessed self-concept, interest and self-efficacy in biology and physics as well as students' epistemic beliefs.[5] Afterwards, participants completed the prior knowledge test. This phase took roughly 30 minutes. During the subsequent learning phase, students were presented with a written introduction to the topic of energy and the topic of experiments. Afterwards, students studied two video modeling examples that explained how to conduct controlled experiments (control of variables strategy) that varied according to condition. Each participant had a headset to listen to the models' comments. The modeling examples stopped several times, at which point students were asked about critical aspects of the examples. After each example, students were instructed to rate their cognitive load. After studying the modeling examples, students solved a training inquiry task. During the complete learning phase, students could consult the experimenter only for technical assistance. On the second day, the second learning phase took place. It was identical to the first learning phase except for the learning content, which was generating hypotheses. In the final lesson on the third day, all participants completed the posttest that contained the achievement tests for scientific reasoning and domain knowledge as well as the two test inquiry tasks.

[5] These variables were assessed only to ensure the comparability of conditions regarding students' entry characteristics and will not be considered any further in the manuscript.

3.2 Results

Table 2 presents descriptive statistics for the pretest data (scientific reasoning test, domain knowledge test) per condition. As a check on randomization, the mean scores on the pretest measures were compared with a MANOVA which – as expected – showed no significant differences between conditions (instructional approach: $F(2, 113) = 2.11$, $p = .13$, $\eta_p^2 = .04$, arrangement: $F < 1$, and instructional approach x arrangement: $F(2, 113) = 2.63$, $p = .08$, $\eta_p^2 = .05$).

Table 2

Mean and standard deviations (in parentheses) for the pretest measures.

	Deductive-mixed (n = 27)	Deductive-blocked (n = 30)	Inductive-mixed (n = 30)	Inductive-blocked (n = 31)
Scientific reasoning test (%)	77.78 (13.25)	75.00 (17.31)	72.78 (18.01)	69.00 (15.57)
Domain knowledge test (%)	68.98 (16.40)	75.83 (16.72)	72.08 (17.27)	64.11 (19.83)

3.2.1 Learning phase

3.2.1.1 Cognitive load. To check whether the groups differed in cognitive load during learning, a 2x2 ANOVA with instructional approach (deductive vs. inductive) and arrangement of modeling examples (mixed vs. blocked) was run for each subjective rating scale (see Table 3). For subjective difficulty, learners in the mixed conditions (M = 2.90, SE = 0.17) found the learning material less difficult than learners in the blocked conditions (M = 3.36, SE = 0.16), $F(1, 114) = 4.00$, $MSE = 1.57$, $p = .048$, $\eta_p^2 = .03$, corresponding to a small effect (Cohen, 1988). There was neither an effect of instructional approach nor an interaction effect between instructional approach and arrangement of modeling examples on subjective difficulty (both $Fs < 1$). Overall, learners rated the difficulty of the learning material rather low. For subjective effort, there were no differences between the groups (all $Fs < 1$). Overall, learners rated their effort as medium.

3.2.1.2 Training inquiry tasks. Because of time constraints in the schools, 30 participants did not reach the training inquiry tasks. After testing that these missing values were independent of group membership (χ^2 (3) = 2.59, $p = .46$), I excluded these participants from

Table 3

Means and standard deviations (in parentheses) for the measures in the learning phase

	Deductive-mixed	Deductive-blocked	Inductive-mixed	Inductive-blocked
Cognitive load				
Subjective difficulty (1-7)	3.05 (1.17)	3.42 (1.49)	2.74 (1.03)	3.30 (1.26)
Subjective effort (1-7)	4.10 (1.44)	4.04 (1.43)	4.05 (1.17)	4.27 (1.32)
Training inquiry tasks				
No. of controlled experiments with hypothesis	11.24 (6.74)	5.30 (6.18)	10.46 (8.95)	6.54 (8.78)
No. of identical experiments	5.53 (10.85)	5.52 (12.78)	3.79 (5.91)	3.63 (6.25)
No. of confounded experiments	0.94 (1.68)	0.78 (1.04)	2.04 (6.04)	1.58 (2.67)

the analysis of the training inquiry tasks. Afterwards, I investigated whether the groups consisting of the remaining 88 participants differed in their performance in the training inquiry tasks. A 2x2 MANOVA with instructional approach (deductive vs. inductive) and arrangement of modeling examples (mixed vs. blocked) was run for the number of controlled experiments with hypotheses, the number of identical experiments, and the number of confounded experiments. There was a main effect for arrangement of modeling examples, $F(3, 82) = 2.77$, $p = .047$, $\eta_p^2 = .09$, with a medium effect size. Follow-up ANOVAs showed that participants in the mixed conditions ($M = 10.85$, $SE = 1.25$) performed more controlled experiments with hypotheses than participants in the blocked conditions ($M = 5.92$, $SE = 1.15$), $F(1, 84) = 8.47$, $MSE = 61.69$, $p = .01$, $\eta_p^2 = .09$. There were no differences in the number of identical experiments or the number of confounded experiments between participants in the mixed and the blocked conditions (both $Fs < 1$). There was neither a main effect for instructional approach nor an interaction between instructional approach and arrangement of modeling examples on performance in the training inquiry tasks (both $Fs < 1$). In sum, participants in the mixed conditions performed more controlled experiments during the learning phase than participants in the blocked conditions, reflecting higher scientific reasoning skills.

3.2.2 Posttest

3.2.2.1 Scientific reasoning test. To test whether students improved in the scientific reasoning test from pretest to posttest, a repeated measures ANOVA with the within-subjects factor time (pretest vs. posttest) and the between-subjects factors instructional approach (deductive vs. inductive) and arrangement of modeling examples (mixed vs. blocked) was run. There was a significant main effect of time, $F(1,114) = 19.00$, $MSE = 86.20$, $p < .001$, $\eta_p^2 = .14$. This corresponds to a large effect. Despite the fact that participants already had high levels of prior knowledge in the pretest, participants in general gained knowledge from pretest ($M = 73.64$, $SE = 1.49$) to posttest ($M = 78.91$, $SE = 1.72$; see Table 2 and Table 4). Moreover, there was a significant main effect of instructional approach $F(1,114) = 4.35$, $MSE = 522.48$, $p = .04$, $\eta_p^2 = .04$, corresponding to a small effect. Aggregated over pretest and posttest, participants in the deductive groups ($M = 79.38$, $SE = 2.14$) performed better than participants in the inductive groups ($M = 73.17$, $SE = 2.07$). There was no significant effect for arrangement of modeling examples, $F(1,114) = 1.79$, $MSE = 522.48$, $p = .18$, $\eta_p^2 = .02$, nor were there any interactions (all $Fs < 1$). In sum, learners in all groups improved on the scientific reasoning test from pretest to posttest.

Table 4

Mean and standard deviations (in parentheses) for the posttest measures

	Deductive-mixed	Deductive-blocked	Inductive-mixed	Inductive-blocked
Scientific reasoning test (%)	84.57 (15.74)	80.19 (18.61)	77.96 (19.76)	72.94 (19.71)
Test inquiry tasks				
No. of controlled experiments with hypothesis	4.72 (2.64)	4.10 (2.18)	6.12 (3.93)	4.07 (3.32)
No. of identical experiments	7.44 (7.51)	4.59 (4.20)	7.60 (5.44)	8.90 (9.74)
No. of confounded experiments	0.72 (0.79)	1.21 (1.26)	1.08 (1.08)	1.97 (2.04)
Domain knowledge test (%)	80.25 (14.23)	76.30 (15.64)	76.67 (18.07)	72.76 (18.56)
Transfer tasks domain knowledge (%)	21.99 (11.48)	22.70 (12.79)	22.50 (14.22)	20.26 (10.91)

3.2.2.2 Test inquiry tasks. I hypothesized that learners in the mixed groups as well as learners in the inductive groups would gain higher quality scientific reasoning skills than participants in the blocked and deductive groups. To test this hypothesis, a 2x2 MANOVA was conducted with instructional approach (deductive vs. inductive) and arrangement of modeling examples (mixed vs. blocked) for the number of controlled experiments with hypotheses, the number of identical experiments, and the number of confounded experiments. I excluded from the analysis ten participants who did not conduct any experiments during the test phase. There was a main effect for arrangement of modeling examples, $F(3, 102) = 3.75$, $p = .01$, $\eta_p^2 = .10$, corresponding to a medium effect. Follow-up ANOVAs revealed that - as in the training inquiry tasks - participants in the mixed conditions performed more controlled experiments with hypotheses ($M = 5.42$, $SE = 0.43$) than participants in the blocked conditions ($M = 4.09$, $SE = 0.40$), $F(1, 104) = 5.08$, $MSE = 9.41$, $p = .03$, $\eta_p^2 = .05$ (small effect). Moreover, participants in the mixed conditions ($M = 0.90$, $SE = 0.20$) performed fewer confounded experiments than participants in the blocked conditions ($M = 1.59$, $SE = 0.18$), $F(1, 104) = 6.43$, $MSE = 1.97$, $p = .01$, $\eta_p^2 = .06$ (medium effect). According to the ANOVA, there were no differences in the number of identical experiments between participants in the mixed and the blocked conditions ($F < 1$). There was neither a significant main effect for instructional approach, $F(3, 102) = 2.02$, $p = .12$, $\eta_p^2 = .06$, nor an interaction between instructional approach and arrangement of modeling examples, $F(3, 102) = 1.97$, $p = .12$, $\eta_p^2 = .06$. In sum, participants in the mixed conditions could transfer and apply their higher scientific reasoning skills from the learning phase to the posttest. This is reflected in the higher number of controlled experiments and the lower number of confounded experiments they conducted in the test inquiry tasks.

3.2.2.3 Domain knowledge test. To test whether students had gained domain knowledge from pretest to posttest, a repeated measures ANOVA with the within-subjects factor time (pretest vs. posttest) and the between-subjects factors instructional approach (deductive vs. inductive) and arrangement of modeling examples (mixed vs. blocked) was run. There was a significant main effect of time, $F(1,114) = 16.57$, $MSE = 138.24$, $p < .001$, $\eta_p^2 = .13$, corresponding to a medium effect. Participants in general gained domain knowledge from pretest ($M = 70.25$, $SE = .16$) to posttest ($M = 76.49$, $SE = 1.55$), despite the fact that domain knowledge in the pretest was quite high. There were no significant main effects for instructional approach, $F(1,114) = 2.00$, $MSE = 455.64$, $p = .16$, $\eta_p^2 = .02$, and arrangement of modeling examples, $F < 1$. Moreover, the interaction between instructional approach and arrangement of modeling examples was not significant, $F(1,114) = 1.76$, $MSE = 455.64$, $p = .19$, $\eta_p^2 = .02$, nor was the interaction between time and instructional approach, $F < 1$, or the interaction between

time and arrangement of modeling examples, $F(1,114) = 1.21$, $MSE = 138.24$, $p = .27$, $\eta_p^2 = .01$. However, there was a significant three-way interaction between time, instructional approach, and arrangement of modeling examples, $F(1,114) = 5.88$, $MSE = 138.24$, $p = .02$, $\eta_p^2 = .05$, corresponding to a small effect. Post-hoc tests revealed that the deductive-mixed group and the inductive-blocked group gained domain knowledge from pretest to posttest ($F(1,114) = 12.39$, $p = .001$, $\eta_p^2 = .10$, and $F(1,114) = 8.38$, $p = .01$, $\eta_p^2 = .06$, corresponding to medium effects), whereas there were no significant knowledge gains for the deductive-blocked and the inductive-mixed groups ($p = .88$ and $p = .13$, respectively). In sum, the deductive-mixed and the inductive-blocked groups experienced significant gains in domain knowledge from pretest to posttest.

3.2.2.4 Transfer tasks domain knowledge. Regarding students' performance in the transfer tasks, a 2x2 ANOVA with instructional approach (deductive vs. inductive) and arrangement of modeling examples (mixed vs. blocked) yielded no significant differences among groups (all $Fs < 1$). Overall, performance on the transfer tasks was rather low compared to the high levels in the domain knowledge test.

3.3 Discussion

The present study investigated the effects of different designs of video modeling examples on scientific reasoning skills and domain knowledge in a simulation-based inquiry environment. First, I varied whether the video modeling examples were presented using a deductive approach (abstract description of scientific reasoning strategy followed by video modeling examples) or an inductive approach (only video modeling examples, from which the abstract scientific reasoning strategy had to be inferred). I expected that the inductive approach would foster the acquisition of scientific reasoning skills (Hypothesis 1a), whereas the deductive approach would foster declarative domain knowledge (Hypothesis 1b). Second, I varied the arrangement of the video modeling examples. Students were either provided with video modeling examples from two subjects (biology and physics) per lesson (mixed arrangement), or they were provided with video modeling examples from only one subject per lesson (blocked arrangement). I hypothesized that the mixed arrangement would foster the acquisition of scientific reasoning skills (Hypothesis 2a), whereas the blocked arrangement would foster domain knowledge (Hypothesis 2b). In addition, I explored possible interactions between the two factors to figure out whether one factor could compensate for the other.

Results showed that across all groups scientific reasoning skills as measured with a multiple-choice achievement test improved. This was indicated by a gain in performance from pretest to posttest. This study found no support for Hypothesis 1a that an inductive approach would be particularly suited to foster scientific reasoning skills. Moreover, there was no main effect of the deductive approach on domain knowledge (Hypothesis 1b). However, as hypothesized, the mixed arrangement of video modeling examples fostered scientific reasoning skills, indicated by more controlled and fewer confounded experiments in inquiry tasks during learning and in the final test phase (Hypothesis 2a). Finally, there was no clear support for Hypothesis 2b that a blocked arrangement would foster domain knowledge. However, there was an interaction effect of instructional approach and arrangement of video modeling examples on domain knowledge, suggesting that the deductive-mixed and the inductive-blocked groups gained domain knowledge from pretest to posttest.

3.3.1 Deductive vs. inductive instructional approach

The first design feature that was targeted in the context of this study was the instructional approach implemented in the video modeling examples. This design feature addressed the question of how to present the abstract principle (here the scientific reasoning strategy) in the examples. In the deductive approach, the scientific reasoning strategy was introduced first in

an abstract way followed by video modeling examples in which the strategy was applied. In the inductive approach, on the other hand, students received only the video modeling examples in which the strategy was embedded and had to infer the principle themselves. In line with previous research, I hypothesized that an inductive approach would foster the application of scientific reasoning skills (Hypothesis 1a), whereas a deductive instructional approach would foster the acquisition of domain knowledge (Hypothesis 1b). Scientific reasoning skills were assessed via a multiple-choice achievement test and inquiry tasks, whereas domain knowledge was assessed via a multiple-choice achievement test and transfer tasks.

I did not find support for Hypothesis 1a that an inductive approach would foster scientific reasoning skills. Results for the multiple-choice scientific reasoning test suggested that all groups improved their scientific reasoning skills. In contrast to Hypothesis 1a, aggregated over pretest and posttest, participants in the deductive groups performed better in the scientific reasoning test than participants in the inductive groups. However, a multiple-choice test might not be a valid measure of skills (see Section 4.5). Furthermore, there were no differences between groups in terms of the number of controlled experiments in training and test inquiry tasks. This result is not in line with previous research, where an inductive approach that included comparing examples fostered the acquisition of skills (Gentner et al., 2003).

One explanation for the results concerning scientific reasoning skills is that it might have been too difficult for the 8th graders to induce the scientific reasoning strategies from the examples. However, students indicated only medium to low levels of subjective difficulty, that is, they perceived the learning material as relatively easy to understand. Another possible explanation is that the implementation quality of both approaches (deductive and inductive) was high enough to work well in both groups. This explanation is in line with an argument brought forward by Renkl (2015), according to which it is not crucial if one provides or lets learners generate the principle (i.e., deductive or inductive approach) for the acquisition of principle-based cognitive skills. Rather, it is important that the implementation quality of the chosen option is high so that, for example, the attention of learners is directed to central concepts through prompts. To ensure high implementation quality, I used prompts in both approaches that aimed to optimize students' learning. Students assigned to the deductive approach had to explain the solution steps of the video modeling examples to themselves, whereas students assigned to the inductive approach had to compare the video modeling examples and find commonalities and differences. It is possible that these prompts were successful in fostering schema construction on how to conduct controlled experiments in both

groups. This explanation is further corroborated by the rather low number of confounded experiments that participants in the deductive approach as well as the inductive approach performed in the posttest (cf. Table 4). Finally, a recent meta-analysis on teaching the control-of-variables strategy found no differences in student performance between studies that explicitly provided a CVS rule (deductive approach) and studies that did not provide a CVS rule (Schwichow et al., 2016).

Analogously, there was no clear support for Hypothesis 1b that a deductive approach would foster domain knowledge. This result is not in line with previous research that showed that deductive approaches that present a rule followed by an example foster the acquisition of declarative knowledge and concepts (Seidel et al., 2013; Tomlinson & Hunt, 1971). Moreover, regarding the performance in the transfer tasks for domain knowledge, there were no performance differences between groups. However, the transfer tasks probably required more than the reproduction of declarative knowledge and concepts that should have been fostered by a deductive approach. This explanation is further corroborated by the modest performance of students on the domain knowledge transfer tasks.

Finally, there were no differences between the groups who learned with the deductive approach and the inductive approach regarding subjective difficulty and subjective efforts of learners.

Taken together, the results suggest that the instructional approach does not seem to be crucial for the acquisition of scientific reasoning skills and domain knowledge if the approach is well implemented. Although the deductive approach had an advantage regarding performance in the scientific reasoning achievement test, this result should be interpreted with caution since the scientific reasoning achievement test might not be a valid measure of scientific reasoning skills (see Section 4.5). Moreover, both approaches in the present study worked equally well in fostering scientific reasoning skills as measured by performance in inquiry tasks. Finally, there was no clear evidence of an advantage of the deductive approach with regard to domain knowledge.

3.3.2 Mixed vs. blocked arrangement

The second design feature addressed the question of how to arrange multiple examples. In a blocked arrangement, one scientific reasoning strategy was taught with examples using the same subject (biology or physics). In a mixed arrangement, in contrast, one scientific reasoning strategy was taught with examples from different subjects (biology and physics). In line with

research by Quilici and Mayer (1996), I hypothesized that a mixed arrangement would foster scientific reasoning skills (Hypothesis 2a), whereas a blocked arrangement would foster domain knowledge (Hypothesis 2b).

Results regarding the scientific reasoning achievement test suggested that there were no differences between groups. However, the question arises as to whether the multiple-choice scientific reasoning achievement test was a valid measure of inquiry skills, since performance results in the inquiry tasks differed. Namely, for the training and test inquiry tasks, I found that participants in the mixed groups who learned one scientific reasoning strategy with video modeling examples from two subjects (biology and physics) acquired higher quality scientific reasoning skills than participants in the blocked conditions. In particular, participants in the mixed conditions were already performing more controlled experiments in the test inquiry tasks during the learning phase. Moreover, they also performed more controlled and fewer confounded experiments in the test inquiry tasks in the posttest. This result confirms Hypothesis 2a and is in line with the research of Quilici and Mayer (1996). Observing one scientific reasoning strategy applied in different contexts – here biology and physics – highlights that the strategy is valid in different contexts. A schema for one scientific reasoning strategy that is encoded with examples from different contexts seems to facilitate the application of this strategy to new contexts. Speaking more generally, the mixed arrangement seems to better foster the transfer of high-quality scientific reasoning skills. In addition, students who learned with a mixed arrangement rated the learning material as less difficult during learning than students who learned with a blocked arrangement. However, there were no differences between the groups in subjective effort.

The results of this study gave no clear support for Hypothesis 2b that a blocked arrangement would foster the acquisition of domain knowledge. There was no main effect of the blocked arrangement on domain knowledge. Moreover, regarding the results of the transfer tasks for domain knowledge, there were no differences between groups. This is not in line with previous research which showed that a blocked arrangement, where one principle is always associated with the same context or surface story, focuses the attention of learners on the context (Quilici & Mayer, 1996). However, in the studies of Quilici and Mayer (1996), the context was irrelevant for learning. Consequently, they did not test whether learners acquired knowledge about the context. The present study, in contrast, targeted declarative domain knowledge in a relevant context. This might explain the difference in findings.

Taken together, the results suggest that the arrangement of video modeling examples influences the acquisition of scientific reasoning skills, at least when measured using inquiry tasks. Students who learned with a mixed arrangement demonstrated meaningful experimentation behavior, reflected in the higher number of controlled experiments and the lower number of confounded experiments in posttest. Additionally, the higher number of controlled experiments in the learning phase reflected that they had already understood the scientific reasoning strategies during the learning process. Finally, the study showed no clear advantage of a blocked arrangement on domain knowledge.

3.3.3 Interaction of instructional approach and arrangement

In contrast to the expected main effects of example arrangement and instructional approach on domain knowledge, there was an interaction between the two factors. A blocked arrangement combined with an inductive instructional approach, as well as a mixed arrangement combined with a deductive instructional approach, fostered the acquisition of domain knowledge from pretest to posttest. One explanation for this interaction relates to the complexity of the factors manipulated in the present experiment. For the instructional approach as well as for the arrangement of video modeling examples, one factor level might have been more complex or more cognitively demanding than the other. The inductive approach, for example, might have been more complex than the deductive approach since students had to infer the scientific reasoning strategy from the examples. Moreover, the mixed arrangement might have been more cognitively demanding than the blocked arrangement since students had to deal with two subjects instead of one. The combination of one more complex and one less complex factor level, that is the deductive-mixed and the inductive-blocked groups, might therefore have resulted in an optimal demand level or germane load for students. However, there were no differences in subjective effort between the groups. Currently, there is no explanation for this interaction effect and further replication of the result pattern is needed. Finally, there were no compensatory interaction effects for the factors regarding scientific reasoning and domain knowledge.

3.3.4 Conclusion

In conclusion, the video modeling examples were found to enhance students' scientific reasoning skills measured with an achievement test and inquiry tasks, as in the study by Mulder et al. (2014). In contrast to Mulder et al. (2014), the video modeling examples of the present study also helped students gain new domain knowledge. Taken together, the results suggest that

it is important to carefully design video modeling examples, since the design might influence what students learn.

4 General discussion

To become responsible citizens in our knowledge-based societies, individuals need skills to understand scientific knowledge. Scientific reasoning skills include the ability to understand the scientific process in different disciplines, and to produce and interpret scientific results (Fischer et al., 2014). Thus, scientific reasoning is not only important for scientists but for everyone in everyday life. Consequently, fostering the acquisition of scientific reasoning skills is considered a main goal of science education (National Research Council, 2012; OECD, 2007).

In general, there are two promising approaches to foster the acquisition of scientific reasoning at schools: inquiry learning, which argues for learning by doing (Hmelo-Silver et al., 2007; Kuhn & Dean, 2005), and example-based learning, which argues for learning by being told (Renkl, 2014; van Gog & Rummel, 2010). Whereas in inquiry learning learners think and act like scientists, for instance, by conducting experiments in the natural sciences, in example-based learning learners study examples showing them how to think and act like scientists. There has been a long debate about the effectiveness of the two teaching philosophies (Hmelo-Silver et al., 2007; Kirschner et al., 2006; Klahr & Nigam, 2005; Kuhn & Dean, 2005; R. E. Mayer, 2004). As both approaches are associated with benefits and drawbacks, the present thesis investigated how to foster students' acquisition of scientific reasoning skills at schools with inquiry and example-based learning.

In the following, the main results of the two studies of the present thesis are summarized. Then, I discuss the results with regard to the theoretical background and present theoretical implications. Subsequently, practical implications are derived. Next, strengths of the present thesis are discussed, followed by limitations and future directions. Finally, I close the general discussion with a concluding statement.

4.1 Summary of main results

The first study addressed the delivery of examples, or more specifically, whether and how video modeling examples and inquiry tasks should be combined to foster scientific reasoning skills. Based on prior research it was not clear whether there were advantages of combining examples and inquiry activities over learning from examples only. In addition, combining the two approaches raised the question of how to sequence examples and inquiry activities. On the one hand, research showed that presenting examples before problems reduced

learners' cognitive load and helped them to create a problem-solving schema (Leppink et al., 2014; Renkl, 2014; van Gog et al., 2011). On the other hand, research showed that solving a problem first made learners aware of knowledge gaps and motivated them to study a subsequent example more closely (Arena & Schwartz, 2014; Loibl & Rummel, 2014). In line with previous research on worked examples (Leppink et al., 2014; Renkl, 2014; van Gog et al., 2011), results indicated an advantage for providing video modeling examples before or instead of but not after an inquiry task. Participants who watched a video modeling example first reported less mental effort in the first training phase and showed better learning outcomes in the first intermediate test as well as in the posttest. Moreover, they were more confident about their future test performance after the second training phase. In contrast to previous research (Baars et al., 2016; Baars, van Gog, et al., 2014), I did not find the result that learners overestimated their performance after studying examples. In my study, all students underestimated their performance.

The second study addressed the design of examples. I examined how the design of video modeling examples could be optimized to foster scientific reasoning skills and domain knowledge simultaneously. I varied two design aspects: the instructional approach and the arrangement of examples with regard to their structural and context features. The instructional approach of an example refers to how the abstract principle of an example is introduced. In a deductive instructional approach, the principle was presented up front, followed by several examples in which the principle was applied. In an inductive instructional approach, in contrast, the abstract principle was not explicitly presented but embedded in the examples and had to be inferred by the learners. Prior research has shown that a deductive instructional approach is especially suited to foster domain knowledge (Seidel et al., 2013; Tomlinson & Hunt, 1971), whereas an inductive instructional approach is especially suited to foster the acquisition of skills (Gentner et al., 2003; Seidel et al., 2013). The second design feature, the arrangement of examples, refers to the question of how to arrange multiple examples with regard to their structural and context features. In the present thesis, the structural features were scientific reasoning strategies and the context features were different subjects (biology and physics). The same scientific reasoning strategy was explained in either a mixed arrangement using examples from different subjects (biology and physics,) or in a blocked arrangement using examples from the same subject (only biology or only physics). Prior research has shown that the blocked arrangement fostered knowledge about the context (domain knowledge), whereas the mixed arrangement fostered the application of the principle (scientific reasoning skills; Quilici & Mayer, 1996). In line with the first study, results showed that video modeling examples

improved scientific reasoning skills from pretest to posttest as measured with a multiple-choice achievement test. In contrast to prior research (Gentner et al., 2003; Seidel et al., 2013; Tomlinson & Hunt, 1971), there were no differential effects of the instructional approach on scientific reasoning skills and domain knowledge. However, in line with Quilici and Mayer (1996), the mixed arrangement of video modeling examples fostered scientific reasoning skills, indicated by more elaborate experimenting behavior during learning and in the posttest. Finally, in contrast to my hypothesis, no effect of the blocked arrangement on domain knowledge emerged. However, there was an interaction effect of instructional approach and arrangement of video modeling examples on domain knowledge, suggesting that the deductive-mixed and the inductive-blocked groups gained domain knowledge from pretest to posttest.

4.2 Theoretical implications

In the following, the results of the two studies are discussed with regard to the theoretical background. After targeting the effectiveness of video modeling examples and their potential drawbacks of inducing illusions of understanding, I explicate possible theoretical implications with regard to the delivery and the design of video modeling examples.

4.2.1 Why are video modeling examples effective in fostering the acquisition of scientific reasoning skills?

Both studies found a positive effect of studying video modeling examples on the acquisition of scientific reasoning skills. In the first study, studying a video modeling example first led to better learning outcomes than solving an inquiry task first. In addition, these better learning outcomes by the example-first groups were achieved with less mental effort. In the second study, performance in the scientific reasoning test improved from pretest to posttest across all groups.

Similar to results from previous research, these results indicate a worked example effect (Cooper & Sweller, 1987; Sweller & Cooper, 1985). That is, studying examples was more beneficial for learning than solving inquiry tasks. This effect can be explained with the different cognitive processes that are evoked by studying examples or solving problems (Sweller et al., 1998). In problem solving, novice learners often must process a large amount of information simultaneously. This induces a high working memory load, which is not effective for learning (Sweller, 1988). Studying examples, on the other hand, reduces cognitive load in learners and helps them to focus their attention on relevant solution steps to build a cognitive problem solving schema (Sweller et al., 1998).

The present thesis extends classic example-based learning research in two ways. First, I used examples to teach scientific reasoning, a cognitive skill that is not highly structured. Traditionally, worked examples have been used to foster performance in highly structured cognitive tasks such as algebra (Cooper & Sweller, 1987; Sweller & Cooper, 1985), statistics (Quilici & Mayer, 1996), geometry (Paas & van Merriënboer, 1994; Schwonke et al., 2009), or physics (Kalyuga et al., 2001; Reisslein et al., 2006). Such highly structured tasks usually can be solved using a straightforward algorithmic solution procedure (Hilbert et al., 2008). Scientific reasoning, however, is not a straightforward algorithmic procedure. It requires taking steps backward and repeating certain actions. Thus, scientific reasoning is an iterative and cyclical process that partly depends on preceding steps (Mulder et al., 2014). Consequently, it

was not clear whether example-based learning would be suited to foster such a less structured skill. Based on these considerations, it is interesting that I found a worked example effect for fostering the acquisition of scientific reasoning skills. Nevertheless, this result is in line with more recent research on example-based learning. This research showed that worked examples can also be effective for less structured cognitive skills such as learning how to apply an instructional systems design methodology (Hoogveld, Paas, & Jochems, 2005), learning argumentation skills (Schworm & Renkl, 2007), or learning to reason about legal cases (Nievelstein, van Gog, van Dijck, & Boshuizen, 2013).

Second, this thesis extends classic example-based learning research in that I used video modeling examples instead of text-based worked examples. Text-based worked examples have been used to teach performance in highly structured cognitive skills with a straightforward algorithmic solution procedure (e.g., Cooper & Sweller, 1987; Paas & Van Merriënboer, 1994; Quilici & Mayer, 1996; Sweller & Cooper, 1985). As scientific reasoning is a less structured skill, text-based worked examples with a straightforward solution procedure did not seem to be suitable. Therefore, I decided to use video modeling examples. In video modeling examples, the model can explain his or her thoughts and heuristics while solving a problem. Accordingly, this approach might be better suited to capture the rationale behind scientific reasoning. However, it was not yet clear, whether video modeling examples would have similar effects to text-based worked examples. The results of the present thesis suggest that they do. Watching a video modeling example reduced cognitive load in learners and enhanced their learning outcomes. This effect has robustly been found in research on text-based worked examples (for a review see Renkl, 2014). Thus, as suggested by a review of van Gog and Rummel (2010), the underlying mechanisms might be similar for both approaches of example-based learning. This is also in line with a study comparing text-based worked examples with video modeling examples, in which the authors found no differences between the different example formats regarding learning, near transfer, effort reduction, self-efficacy, and perceived competence (Hoogerheide, Loyens, & van Gog, 2014).

In line with previous research, I can conclude that studying examples, whether text-based worked examples or video modeling examples, is an effective way of enhancing cognitive skill acquisition for highly structured but also for less structured tasks such as scientific reasoning (Renkl, 2014, 2015; van Gog & Rummel, 2010).

4.2.2 Why are video modeling examples not associated with the drawback of inducing illusions of understanding?

This thesis brings together the notion of monitoring with more complex learning, namely learning to solve problems. Judgments of learning or monitoring accuracy, in general, are important aspects of self-regulated learning, which becomes more and more important in our knowledge-based societies. Learners must be able to accurately judge what they have learned in order to control their learning process (Dunlosky & Rawson, 2012). Overconfidence, for instance, might have detrimental effects on learning as learners might terminate studying before they have learned everything (Dunlosky & Rawson, 2012). However, judgments of learning have predominantly been investigated using word pairs or expository texts as learning materials (e.g., Dunlosky & Rawson, 2012; Eitel, 2016; Thiede & Anderson, 2003; Thiede et al., 2003; van Loon et al., 2014). Only a few recent studies investigated judgments of learning when learning to solve problems (Baars et al., 2016; Baars, Visser, van Gog, de Bruin, & Paas, 2013; Baars, van Gog, et al., 2014; Baars, Vink, van Gog, de Bruin, & Paas, 2014). These studies already suggest that results from studies using word pairs or expository text cannot necessarily be transferred to learning to solve problems. Delaying judgments of learning, for instance, enhances monitoring accuracy in learning word pairs, but appears to have hardly any effect on the accuracy of judgments of learning in problem solving (Baars et al., 2016; Baars, van Gog, et al., 2014).

Baars and colleagues also found that studying examples in learning to solve problems induced overconfidence or illusions of understanding in learners (Baars et al., 2016; Baars, van Gog, et al., 2014). That is, learners think they understood everything when they actually did not. Illusions of understanding might be even more likely to occur when using video modeling examples. Dynamic visualizations like videos are commonly associated with entertainment. Therefore, students may underestimate the effort necessary to understand what is being conveyed through a dynamic visualization (underwhelming effect; Lowe, 2004). Consequently, I expected that video modeling examples would result in overconfident judgments of learning. However, video modeling examples did not induce overconfident but underconfident judgments.

There might be different explanations for this unexpected result. First, underconfidence might be a result of scientific reasoning as the learning content. Scientific reasoning is closely related to epistemological beliefs about knowledge and knowing in science (J. Mayer, 2007). One dimension of epistemological beliefs is the certainty of knowledge. Sophisticated

epistemological beliefs with regard to the certainty of knowledge reflect the belief that there might be more than one answer to complex problems and that scientists often disagree about what is true in science (Conley et al., 2004). If students in my study held sophisticated beliefs and thought that it is hard to answer questions about scientific experiments in general, this could have resulted in underestimation of their own performance.

Second, the underconfidence could be a result of an underconfidence-with-practice effect (Koriat et al., 2002) with learners relying on a memory-for-past-test heuristic (Finn & Metcalfe, 2007). This effect occurs when there are repeated study-test cycles as in the first study of the present thesis. When giving their judgment of learning after the second training phase, learners might have based their judgment on their test performance after the first training phase. When learners remembered that they were not able to answer a given item on this test, they might have judged it unlikely to answer it in a future test. However, they disregarded the fact that there was a second training phase, which probably enhanced their future test performance.

Finally, the underconfidence could be attributed to the fact that we used video modeling examples rather than text-based worked examples. I expected that video modeling examples might be even more likely to induce overconfidence or illusions of understanding in learners as videos might have been associated with leisure time rather than with learning (Lowe, 2004). However, results suggest that video modeling examples did not induce illusions of understanding in learners. In contrast, studying video modeling examples resulted in underestimation of future test performance. Thus, I was able to prevent a possible underwhelming effect caused by using a dynamic visualization. The models that were utilized might offer a suitable explanation for this circumstance. First, I used authentic models of the same age as students. Second, I used coping models that made errors that they then corrected (van Gog & Rummel, 2010). The use of authentic models might have helped students to identify with the models (Schunk, 1987). Observing these authentic models encountering difficulties in solving an inquiry task might have prevented students from being overconfident. This might explain the different results compared to the studies by Baars et al. (2016) and Baars, van Gog, et al. (2014) using text-based worked examples. Thus, video modeling examples in the present thesis were not associated with the drawback of inducing illusions of understanding. Future studies should investigate if this holds also true for video modeling examples conveying different learning content.

4.2.3 How should video modeling examples be delivered to optimize their effectiveness?

One research question of the present thesis was if and how video modeling examples should be combined with inquiry tasks. It was not clear, whether there were advantages of combining examples and problems over studying examples only. Sweller and Cooper (1985), for instance, proposed that coupling examples with problems might be more motivating than studying examples only. Combining video modeling examples and inquiry tasks furthermore raised the question of whether the sequence of examples and problems had an effect on learning outcomes. Example-problem pairs might be advantageous because studying an example first could reduce cognitive load in learners and help them to build a problem-solving schema, which could then be applied during solving the subsequent practice problem (van Gog et al., 2011). However, problem-example pairs might also be advantageous because solving a problem first might enable learners to become aware of what they already have or have not learned and thus focus their attention on the respective knowledge gaps in the subsequent example (Arena & Schwartz, 2014; Loibl & Rummel, 2014).

Results showed that learners who studied a video modeling example followed by an inquiry task did not differ from learners who studied two video modeling examples, regarding cognitive load and learning outcomes. This result was in line with previous research using text-based worked examples (Leppink et al., 2014; van Gog et al., 2011). Consequently, results indicate that there are no advantages of combining video modeling examples and inquiry tasks over studying video modeling examples only. However, due to the limited number of examples and problems used in the first study, this conclusion should be interpreted with caution. In longer training phases with a higher number of examples and tasks there might be motivational advantages of the example-task group compared to the example-example group. Previous studies have argued that alternating examples and tasks might be more motivating than only studying examples (Sweller & Cooper, 1985). Trafton and Reiser (1993), for instance, have found advantages of example-task pairs over a condition in which learners first studied a sequence of examples followed by a sequence of problems. However, Trafton and Reiser (1993) used short text-based worked examples. The modeling examples in the present thesis were more comprehensive and video-based. It is possible that the video modeling examples were thus more motivating than short text-based examples. Consequently, whether findings by Trafton and Reiser (1993) also hold true for video modeling examples should be subject to future studies.

Nevertheless, if examples and problems are combined, the sequence of examples and problems has to be taken into account. In line with previous research (Leppink et al., 2014; van Gog et al., 2011), I found that examples should be presented first followed by problems. This sequence led to lower cognitive load and better learning outcomes than problems followed by examples. Thus, it seems to be crucial to present a video modeling example first. Afterwards, solving an inquiry task or watching a second video modeling example seem to be equally effective and efficient to foster scientific reasoning skills.

4.2.4 How should video modeling examples be designed to optimize their effectiveness?

Apart from the delivery of examples, another important aspect influencing the effectiveness of example-based learning is the design of examples (Atkinson et al., 2000; van Gog & Rummel, 2010). The design aspects addressed in the present thesis were the instructional approach and the arrangement of examples. Each aspect had two facets that were expected to foster either domain knowledge or scientific reasoning skills.

I did not find the expected differential effects of the instructional approach on scientific reasoning skills and domain knowledge. This result might be explained by the high implementation quality of both approaches (Renkl, 2015). Learners in both approaches received prompts that seemed to have helped them to concentrate on central concepts. This result is also in line with a recent meta-analysis on teaching the control-of-variables strategy (Schwichow et al., 2016). The authors did not find differences in student performance between studies that explicitly provided a CVS rule (deductive approach) and studies that did not provide a CVS rule (Schwichow et al., 2016). Hence, the instructional approach does not seem to be crucial for the acquisition of scientific reasoning skills and domain knowledge if the approach is well implemented.

As expected, there was an advantage of the mixed arrangement on the acquisition of scientific reasoning skills. That is, learners who learned one scientific reasoning strategy with examples from different subjects were better able to conduct controlled experiments in test inquiry tasks. This was in line with previous research that showed that a mixed arrangement focused the attention of learners on the underlying structural features of an example, which is a necessary prerequisite for skill acquisition (Quilici & Mayer, 1996). There was no effect of the blocked arrangement on domain knowledge, a result that stood in contrast to findings by Quilici and Mayer (1996), who showed that a blocked arrangement focused learners' attention on the context. However, the context in the study of Quilici and Mayer (1996) was irrelevant for

learning, whereas in my study it was relevant for learning. It is possible that learners focused their attention on the context but at the same time, they did not process it deeply. Nevertheless, there was an interaction effect between the instructional approach and the arrangement on domain knowledge. Learners in the deductive-mixed and the inductive-blocked group acquired domain knowledge from pretest to posttest. Currently, there is no explanation for this interaction effect and further replication of the result pattern is needed.

In conclusion, the design of video modeling examples does influence the acquisition of scientific reasoning skills and domain knowledge. If the aim is to foster both skills and knowledge; a deductive instructional approach together with a mixed arrangement is to be preferred.

4.3 Practical implications

From a practical perspective, video modeling examples seem to be well suited to foster the complex and cyclical skills involved in scientific reasoning. Thus, the use of video modeling examples can be recommended to educators, at least when their learners have low prior knowledge. Video modeling examples might be especially helpful in the beginning of learning. It is not yet completely clear, whether there are benefits of combining video modeling examples with inquiry problems. However, there do not seem to be disadvantages either, at least when examples are presented first followed by problems.

In addition, if the aim of educators is to teach scientific reasoning skills and domain knowledge simultaneously, which might often be the case in educational practice, the examples should be carefully designed. Introducing the scientific reasoning principle in an abstract way, followed by several examples from different contexts or subjects (deductive approach and mixed arrangement), might be best suited for fostering scientific reasoning skills and domain knowledge simultaneously. Thus, when educators create their own video modeling examples it might be beneficial if educators from several disciplines work together.

Educators can use video modeling examples to augment their in-class teaching or as materials for learners to study on their own at home or during self-regulated learning sessions in school. However, the production of good video modeling examples requires some planning and time. Educators have to think about the best way of explaining a scientific reasoning strategy, look for appropriate simulated experiments, take a screen capture of an experiment and edit the resulting video according to guidelines for the design of instructional videos (e.g., van der Meij & van der Meij, 2013). This might be too effortful for a single use of video modeling examples. However, once a good video modeling example has been produced it can easily be reused and shared with others. Additionally, video modeling examples can be integrated in learning software or e-books to be used by a larger audience than the class of a single teacher. Thus, the production of video modeling examples could also be interesting for commercial providers such as schoolbook publishers.

4.4 Strengths of the present thesis

The present thesis is associated with multiple strengths with regard to the quality of the studies and to theoretical aspects.

One important strength of this thesis is the operationalization of scientific reasoning skills. Scientific reasoning skills were assessed with two different measures. First, I used an established achievement test to assess scientific reasoning skills. This test was designed to measure the ability to apply knowledge about experimental scientific practices (Koenen, 2014). An achievement test is an efficient way of assessing scientific reasoning skills, but it might not be a valid instrument for measuring skills. An achievement test does not capture behavior directly but rather knowledge about behavior or skills. Thus, an achievement test does not require real actions from students. Therefore it is likely that such a test assesses so-called inert knowledge, that is, knowledge that can be reproduced in assessment situations but that would not be spontaneously applied to real life problem-solving situations (Renkl et al., 1996). To complement the achievement test with a behavioral measure, I also used inquiry tasks as a second measure for scientific reasoning skills. For this purpose, I video-recorded learners' experimentation behavior while they were solving virtual inquiry tasks and analyzed their behavior with regard to important scientific reasoning skills. This assessment is very close to actual scientific reasoning and thus less likely to only capture rote understanding of science (DeBoer et al., 2008; Gobert et al., 2013; Quellmalz et al., 2013).

Another strength of the present thesis is the learning material that was utilized. The virtual experiments were carefully chosen within the context of the crosscutting concept of energy; moreover, the domain-specific contents were developed with the help of experts in subject matter didactics. The topic of energy is part of the course curriculum for students in Grades 7 and 8. Thus, the learning material was relevant to the students. This is in contrast to earlier studies that investigated scientific reasoning using knowledge-lean tasks to avoid the influence of prior knowledge (e.g., Kuhn & Phelps, 1982; Siegler & Liebert, 1975). The latter approach might be critiqued as yielding an artificial learning situation, as in real-life settings learners will usually have some prior knowledge. Thus, to be able to draw any conclusions relevant for educational practice it appears to be more appropriate to use realistic learning materials. In addition, I used the same material in both studies. Therefore, it is easier to integrate the results of both studies.

Another strength of the present thesis is that both studies were conducted in real school settings. In contrast to controlled lab studies, there are many confounding variables in real school settings. For instance, there might be more noise and distraction by seatmates so that it is harder for students to concentrate on the learning material. Thus, field studies have a higher ecological validity. That is, results found in a field study can be more easily generalized to real-life settings. In addition, classroom studies are more likely to influence the praxis of teaching and thus have a larger impact relative to laboratory studies (Hofstein & Lunetta, 2004).

In addition, the present thesis corroborates and extends previous research on inquiry learning vs. direct instruction as well as on example-based learning. First, the results demonstrate that examples as one form of direct instruction were more effective than inquiry learning for fostering scientific reasoning skills. This result is line with the meta-analysis of Alfieri et al. (2011) indicating that inquiry learning with no or minimal guidance was less beneficial for learning than direct instruction. Moreover, it is also in line with the result that the effectiveness of worked examples was not different from guided inquiry learning (Alfieri et al., 2011). In general, the results of the present thesis speak for the information-processing approach advocating for direct instruction or learning by being told (Kirschner et al., 2006; Klahr & Nigam, 2005). Second, the present thesis extends classic example-based learning research (e.g., Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Cooper & Sweller, 1987; Sweller, 1988; Sweller & Cooper, 1985) to the more recent trend of using video modeling examples (Hoogerheide et al., 2014; Hoogerheide, Loyens, & van Gog, 2016; Hoogerheide, van Wermeskerken, Loyens, & van Gog, 2016). Results from traditional worked example research can partially be transferred to video modeling examples. There seems to be a worked example effect for video modeling examples as shown by reduced cognitive load and better learning outcomes through studying examples. However, contrary to worked examples (Baars et al., 2016; Baars, van Gog, et al., 2014), video modeling examples in the present thesis did not induce illusions of understanding. As described above in Section 4.2.2, the use of authentic coping models might have raised learners' awareness for the complex process of scientific reasoning, which might have prevented illusions of understanding. Future research should investigate if this advantage of video modeling examples can be replicated with different learning materials.

Finally, this thesis showed that students' scientific reasoning skills could be fostered with a relatively short training program that targeted central aspects of scientific reasoning (control-of-variables strategy and hypothesis generation). This is remarkable, considering that some inquiry-based trainings covered several months (Dean & Kuhn, 2007). In schools,

however, teaching scientific reasoning skills is only one learning goal among others. Thus, spending several months for reaching one learning goal is inefficient. In contrast to that, the training program of the present thesis enhanced the development of scientific reasoning skills within two to three lessons. Similarly, a recent meta-analysis found that the duration of a study did not affect the effectiveness of guided inquiry on scientific reasoning skills (Lazonder & Harmsen, 2016). In addition, the results of the present thesis are in line with another recent meta-analysis showing that teaching the control-of-variables strategy is possible and can be effective (Schwichow et al., 2016).

4.5 Limitations and future directions

Despite the strength of the present thesis, there are also some limitations and open questions that need to be considered. First, the operationalization of scientific reasoning in the present thesis is a strength and a limitation at the same time. With the inquiry tasks, I measured two specific strategies (control-of-variables strategy, generation of hypotheses) that can be considered to be domain-general (C. Zimmerman, 2007). Scientific reasoning skills in general also include other aspects such as the evaluation of results (Klahr & Dunbar, 1988) or in some definitions the communication of results (Fischer et al., 2014; Kuhn, 2010). In addition, there might also be domain-specific aspects of scientific reasoning skills. For instance, disciplines may vary in what is regarded as acceptable evidence (Fischer et al., 2014). Thus, the operationalization of scientific reasoning skills used in the studies of the present thesis only captured certain aspects of scientific reasoning skills. Future studies should also include other scientific reasoning strategies or even the complete inquiry cycle as well as more domain-specific aspects to investigate whether the results of the present thesis can be generalized.

Another possible limitation of the present thesis pertains to learner prerequisites. As described in the theoretical background (Section 1.2.2.2), learner prerequisites are an important factor influencing the effectiveness of example-based learning (van Gog & Rummel, 2010). The most important learner characteristic is learners' prior knowledge. Example-based learning seems to be especially beneficial for novices in a domain because studying examples seems to help them to construct cognitive problem-solving schemata for future problem-solving situations (Renkl, 2014). Previous research has shown that with increasing domain knowledge there is an expertise-reversal-effect (Kalyuga et al., 2001). That is, with increasing expertise the advantage of studying examples first might decrease or even vanish. For learners with high prior knowledge it might even be beneficial to solve inquiry tasks only (Kalyuga, 2007; Kalyuga, Ayres, Chandler, & Sweller, 2003). In the present thesis, students with low prior domain knowledge regarding the topic of energy in biology and physics learned with the training program. It is possible that the selection of learners with low prior knowledge affected at least the results of the first study. Specifically, in the first study studying examples initially led to lower cognitive load and better learning outcomes than solving an inquiry task first. Thus, future research should also include learners that are more experienced and investigate potential moderating effects of prior knowledge. However, even if there was a moderating effect of prior knowledge, the use of virtual experiments and video modeling examples offers the possibility for adaptive training programs. Thus, in a training program covering, for instance, a complete

teaching unit rather than single lessons the program could be adapted to learners' prior knowledge. When learners have low prior knowledge, they could provided with video modeling examples that show how to solve inquiry problems. With increasing expertise, learners might be confronted with completion problems with increasingly more steps for the learners to complete. Finally, the examples could be completely replaced by practice problems. This fading strategy has been proven effective in other contexts (Atkinson, Renkl, & Merrill, 2003; Renkl, Atkinson, Maier, & Staley, 2002).

Furthermore, a limitation of the present thesis might be that the mechanisms underlying the effect of the video modeling examples remain unexplained. The first study showed that, compared to solving an inquiry task, studying an example reduced cognitive load in learners. Based on previous research, I assumed that this reduced cognitive load helped learners in building a problem-solving schema for future inquiry tasks (Renkl, 2014; van Gog & Rummel, 2010). However, this assumption was not directly tested. Understanding the mechanisms behind video modeling examples might help to further optimize the examples. Therefore, future research should include a measure of the acquired problem-solving schema. For instance, learners could be asked to describe the general approach for testing a hypothesis. Subsequently, their answers might be analyzed as a measure for the acquired problem-solving schema.

Finally, I used video modeling examples in the present thesis assuming that they might be a more valid way to convey the complex cyclical nature of scientific reasoning skills (Hilbert et al., 2008; Mulder et al., 2014). However, I did not test this assumption. Thus, it is possible that text-based worked examples might be equally suited to foster the acquisition of scientific reasoning skills (cf., Hoogerheide et al., 2014). Consequently, it could be interesting to compare the effects of text-based worked examples and video modeling examples for fostering scientific reasoning skills in future studies.

4.6 Conclusion

As the acquisition of scientific reasoning is considered a main goal of science education (National Research Council, 2012; OECD, 2007), the present thesis aimed at developing a digital training program including inquiry tasks and video modeling examples to foster the acquisition of scientific reasoning skills. Across both studies of this thesis, studying video modeling examples followed by solving inquiry tasks was shown to be an effective method to foster the acquisition of scientific reasoning skills and domain knowledge. Thus, learning with the training program developed in the present thesis can support students in acquiring a thinking style that helps them to understand the scientific process across disciplines, to evaluate the validity of scientific claims, to assess the relevance of scientific results, and to apply scientific concepts and methods to generate new knowledge (Fischer et al., 2014).

5 Summary

To become responsible citizens in our knowledge-based societies, individuals need skills to understand scientific knowledge. *Scientific reasoning* skills include the ability to understand the scientific process in different disciplines and to produce and interpret scientific results. Thus, scientific reasoning is not only important for scientists but for everyone in everyday life. Consequently, fostering the acquisition of scientific reasoning skills is considered a main goal of science education.

In general, there are two promising approaches to foster the acquisition of scientific reasoning in schools: *inquiry learning*, which argues for learning by doing and *example-based learning*, which argues for learning by being told. In inquiry learning, learners think and act like scientists, for instance, by conducting experiments in the natural sciences. Conversely, in example-based learning, learners study examples showing them how to think and act like scientists. There has been a long debate about the effectiveness of the two teaching philosophies. As both approaches are associated with benefits and drawbacks, the present thesis investigated how to foster students' acquisition of scientific reasoning skills at schools combining inquiry and example-based learning.

The first study addressed the delivery of examples, or more specifically, if and how video modeling examples and inquiry tasks should be combined to foster scientific reasoning skills. Based on prior research it was not clear whether there were advantages of combining examples and inquiry activities over learning from examples only. In addition, combining both approaches raised the question of how to sequence examples and inquiry activities. Results indicated an advantage for providing video modeling examples before or instead of but not after an inquiry task. Participants who watched a video modeling example first reported less mental effort and exhibited better learning outcomes. Contradicting hypothesized concerns, studying examples was not associated with the drawback of inducing illusions of understanding.

The second study addressed the design of examples, or more specifically, how the design of video modeling examples could be optimized to foster scientific reasoning skills and domain knowledge simultaneously. Results showed that studying video modeling examples improved students' scientific reasoning skills. Introducing the scientific reasoning strategy explicitly or embedding it in the examples had no differential effects on scientific reasoning skills and domain knowledge. However, as hypothesized, teaching one scientific reasoning

strategy using video modeling examples from different subjects rather than from a single subject fostered the acquisition of scientific reasoning skills. In addition, a combination of introducing the scientific reasoning strategy explicitly in combination with video modeling examples from different subjects fostered the acquisition of domain knowledge.

In conclusion, video modeling examples were shown to be an effective way of enhancing students' scientific reasoning skills when being presented before or instead of inquiry tasks. In addition, if the aim of educators is to teach scientific reasoning skills and domain knowledge simultaneously, scientific reasoning strategies should be introduced first followed by examples from different subjects.

6 Zusammenfassung

Mündige Bürgerinnen und Bürger unserer Wissensgesellschaft zu sein, erfordert die Fähigkeit, wissenschaftliche Erkenntnisse zu verstehen und nutzen zu können. *Wissenschaftliches Denken* umfasst, den wissenschaftlichen Prozess in verschiedenen Disziplinen zu verstehen und zu wissen, wie wissenschaftliche Erkenntnisse gewonnen und interpretiert werden können. Wissenschaftliches Denken ist somit nicht nur relevant für Wissenschaftlerinnen und Wissenschaftler, sondern für alle Menschen im täglichen Leben. Deshalb ist die Förderung wissenschaftlichen Denkens ein zentrales Ziel der naturwissenschaftlichen Schulbildung.

Es gibt zwei vielversprechende Ansätze, um wissenschaftliches Denken in der Schule zu fördern: forschendes Lernen und beispielbasiertes Lernen. Beim forschenden Lernen denken und handeln Schülerinnen und Schüler wie Wissenschaftlerinnen und Wissenschaftler, indem sie zum Beispiel selbst naturwissenschaftliche Experimente durchführen. Im Gegensatz dazu studieren Schülerinnen und Schüler beim beispielbasierten Lernen Beispiele, in denen gezeigt wird, wie Wissenschaftlerinnen und Wissenschaftler denken und handeln. Schon lange wird über die Wirksamkeit der beiden Ansätze diskutiert. Da beide Ansätze Vor- und Nachteile haben, untersuchte die vorliegende Dissertation, wie das wissenschaftliche Denken mit forschendem und beispielbasiertem Lernen in der Schule gefördert werden kann. Dazu wurde ein digitales Lernprogramm entwickelt, das sowohl Experimentieraufgaben zu virtuellen Experimenten (forschendes Lernen) als auch videobasierte Lösungsbeispiele zum Experimentieren (beispielbasiertes Lernen) enthielt.

Die erste Studie der vorliegenden Dissertation beschäftige sich mit der Darbietung von Beispielen. Es wurde untersucht, ob und wie videobasierte Lösungsbeispiele mit Experimentieraufgaben kombiniert werden sollten, um wissenschaftliches Denken zu fördern. Auf Grundlage früherer Forschung war unklar, ob eine Kombination Vorteile gegenüber rein beispielbasiertem Lernen bieten würde. Zusätzlich stellte sich durch die Kombination der beiden Ansätze die Frage, in welcher Reihenfolge videobasierte Lösungsbeispiele und Experimentieraufgaben am besten kombiniert werden sollten. Die Ergebnisse der ersten Studie zeigten, dass es vorteilhaft war, Beispiele vor oder statt einer Experimentieraufgaben zu studieren. Ein Beispiel nach einer Experimentieraufgabe zu studieren, brachte dagegen keinen Vorteil. Lerner, die zuerst ein videobasiertes Lösungsbeispiel studierten, erzielten bessere Lernergebnisse, die sie zudem mit weniger Anstrengung erreichten. Entgegen meiner

Hypothese erzeugten videobasierte Lösungsbeispiele keine Illusion des Verstehens. Somit scheinen sie gut geeignet, um wissenschaftliches Denken bei Schülerinnen und Schülern zu fördern.

Die zweite Studie der vorliegenden Dissertation beschäftigte sich deshalb mit der Gestaltung von videobasierten Lösungsbeispielen. Es wurde untersucht, wie videobasierte Lösungsbeispiele gestaltet sein müssen, um wissenschaftliches Denken und gleichzeitig den Erwerb von Fachwissen zu fördern. Auch in der zweiten Studie förderten die videobasierten Lösungsbeispiele das wissenschaftliche Denken von Schülerinnen und Schülern. Ob eine Strategie wissenschaftlichen Denkens zunächst abstrakt eingeführt wurde oder direkt in konkrete videobasierte Lösungsbeispiele eingebettet war, hatte keine differentiellen Effekte auf die Förderung des wissenschaftlichen Denkens und den Erwerb von Fachwissen. Wie vermutet zeigte sich jedoch, dass es vorteilhaft für die Förderung des wissenschaftlichen Denkens war, eine Strategie wissenschaftlichen Denkens mit Beispielen aus unterschiedlichen Fächern zu vermitteln. Zusätzlich förderte eine abstrakte Einführung einer Strategie wissenschaftlichen Denkens kombiniert mit videobasierten Lösungsbeispielen aus unterschiedlichen Fächern den Erwerb von Fachwissen.

Zusammenfassend zeigte sich in der vorliegenden Dissertation, dass videobasierte Lösungsbeispiele eine effektive Methode sind, um das wissenschaftliche Denken bei Schülerinnen und Schülern zu fördern, wenn die Beispiele vor oder statt einer Experimentieraufgabe gezeigt werden. Wenn Lehrer multiple Ziele im Unterricht verfolgen wie die gleichzeitige Förderung von wissenschaftlichem Denken und Fachwissen, sollten sie zunächst die Strategie wissenschaftlichen Denkens abstrakt einführen und diese anschließend mit Beispielen aus unterschiedlichen Fächern darstellen.

7 References

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, *103*, 1–18. http://doi.org/10.1037/a0021017

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of Learning Sciences*, *4*, 167–207. http://doi.org/10.1207/s15327809jls0402_2

Arena, D. A., & Schwartz, D. L. (2014). Experience and explanation: Using videogames to prepare students for formal instruction in statistics. *Journal of Science Education and Technology*, *23*, 538–548. http://doi.org/10.1007/s10956-013-9483-3

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, *70*, 181–214. http://doi.org/10.3102/00346543070002181

Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. *Journal of Educational Psychology*, *95*, 774–783. http://doi.org/10.1037/0022-0663.95.4.774

Baars, M., van Gog, T., de Bruin, A., & Paas, F. G. W. C. (2014). Effects of problem solving after worked example study on primary school children's monitoring accuracy. *Applied Cognitive Psychology*, *28*, 382–391. http://doi.org/10.1002/acp.3008

Baars, M., van Gog, T., de Bruin, A., & Paas, F. G. W. C. (2016). Effects of problem solving after worked example study on secondary school children's monitoring accuracy. *Educational Psychology*. (Advance online publication). http://doi.org/10.1080/01443410.2016.1150419

Baars, M., Vink, S., van Gog, T., de Bruin, A., & Paas, F. G. W. C. (2014). Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. *Learning and Instruction*, *33*, 92–107. http://doi.org/10.1016/j.learninstruc.2014.04.004

Baars, M., Visser, S., van Gog, T., de Bruin, A., & Paas, F. G. W. C. (2013). Completion of partially worked-out examples as a generation strategy for improving monitoring accuracy. *Contemporary Educational Psychology*, *38*, 395–406. http://doi.org/10.1016/j.cedpsych.2013.09.001

Barzilai, S., & Blau, I. (2014). Scaffolding game-based learning: Impact on learning

achievements, perceived learning, and game experiences. *Computers and Education, 70,* 65–79. http://doi.org/10.1016/j.compedu.2013.08.003

Blair, G. M. M. (1940). The validity of the Noll test of scientific thinking. *Journal of Educational Psychology, 31,* 53–59. http://doi.org/10.1037/h0055436

Bortz, J. (1984). *Lehrbuch der empirischen Forschung*. Berlin, Heidelberg: Springer. http://doi.org/10.1007/978-3-662-00468-5

Braaksma, M. A. H., Rijlaarsdam, G., & van den Bergh, H. (2002). Observational learning and the effects of model-observer similarity. *Journal of Educational Psychology, 94,* 405–415. http://doi.org/10.1037/0022-0663.94.2.405

Bruner, J. (1961). The act of discovery. *Harvard Educational Review, 31,* 21–32.

Bybee, R. W. (1997). Toward an understanding of scientific literacy. In W. Gräber & C. Bolte (Eds.), *Scientific Literacy: An International Symposium* (pp. 37–68). Kiel: Institut für Pädagogik der Naturwissenschaften (IPN).

Camtasia Studio. (Version 8.5). [Computer software]. Okemos, Michigan: TechSmith Corporation.

Chang, H., Chen, C., Guo, G., Cheng, Y., Lin, C., & Jen, T. (2011). The development of a competence scale for learning science: Inquiry and communication. *International Journal of Science and Mathematics Education, 9,* 1213–1233. http://doi.org/10.1007/s10763-010-9256-x

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70,* 1098–1120. http://doi.org/10.1111/1467-8624.00081

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). How students study and use examples in learning to solve problems. *Cognitive Science, 13,* 145–182.

Chinn, C., & Malhotra, B. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education, 86,* 175–218. http://doi.org/10.1002/sce.10001

Cierniak, G., Scheiter, K., & Gerjets, P. (2009). Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Computers in Human Behavior, 25,* 315–324. http://doi.org/10.1016/j.chb.2008.12.020

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Collins, A., Brown, J., & Newman, S. (1989). Cognitive apprenticeship: Teaching the craft of

reading, writing, and mathematic. In L. B. Resnick (Ed.), *Knowing, learning, and instruction* (pp. 453–494). Hillsdale, NJ: Erlbaum.

Conley, A. M., Pintrich, P. R., Vekiri, I., & Harrison, D. (2004). Changes in epistemological beliefs in elementary science students. *Contemporary Educational Psychology*, *29*, 186–204. http://doi.org/10.1016/j.cedpsych.2004.01.004

Cooper, G. A., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology*, *79*, 347–362. http://doi.org/10.1037/0022-0663.79.4.347

Davis, E. A. (2000). Scaffolding students' knowledge integration: Prompts for reflection in KIE. *International Journal of Science Education*, *22*, 819–837. http://doi.org/10.1080/095006900412293

de Jong, T., & Lazonder, A. W. (2014). The guided discovery learning principle in multimedia learning. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (2nd ed., pp. 371–390). Cambridge: Cambridge University Press. Retrieved from http://ebooks.cambridge.org/ref/id/CBO9781139547369

de Jong, T., Linn, M. C., & Zacharia, Z. C. (2013). Physical and virtual laboratories in science and engineering education. *Science*, *340*, 305–308. http://doi.org/10.1126/science.1230579

de Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, *68*, 179–201. http://doi.org/10.3102/00346543068002179

Dean, D., & Kuhn, D. (2007). Direct instruction vs. discovery: The long view. *Science Education*, *91*, 384–397. http://doi.org/10.1002/sce.20194

DeBoer, G. E., Abell, C. H., Regan, T., & Wilson, P. (2008). Assessment linked to science learning goals: Probing student thinking through assessment. In J. Coffey, R. Douglas, & C. Stearns (Eds.), *Assessing student learning: Perspectives from research and practice* (pp. 231–252). Arlingtion, VA: National Science Teachers Association.

Decker, P. J. (1980). Effects of symbolic coding and rehearsal in behavior-modeling training. *The Journal of Applied Psychology*, *65*, 627–634. http://doi.org/10.1037/0021-9010.65.6.627

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*, 271–280. http://doi.org/10.1016/j.learninstruc.2011.08.003

Eitel, A. (2016). How repeated studying and testing affects multimedia learning: Evidence for

adaptation to task demands. *Learning and Instruction*, *41*, 70–84. http://doi.org/10.1016/j.learninstruc.2015.10.003

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *33*, 238–244. http://doi.org/10.1037/0278-7393.33.1.238

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., … Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research*, *4*, 28–45. http://doi.org/http://dx.doi.org/10.14786/flr.v2i2.96

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, *95*, 393–408. http://doi.org/10.1037/0022-0663.95.2.393

Gizmos. [Computer simulations]. (2016). Charlottesville: ExploreLearning.

Glaser, R., Schauble, L., Raghavan, K., & Zeitz, C. (1992). Scientific reasoning across different domains. In E. de Corte, M. Linn, H. Mandl, & L. Verschaffel (Eds.), *Computer-Based Learning Environments and Problem Solving* (pp. 345–371). Berlin: Springer.

Glug, I. (2009). *Entwicklung und Validierung eines Multiple-Choice-Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung*. (Nicht veröffentlichte Doktorarbeit). Christian-Albrechts-Universität zu Kiel.

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, *22*, 521–563. http://doi.org/10.1080/10508406.2013.837391

Guzdial, M. (1994). Software-realized scaffolding to facilitate programming for science learning. *Interactive Learning Environments*, *4*, 1–44. http://doi.org/10.1080/1049482940040101

Hardy, I., Kleickmann, T., & Koerber, S. (2010). Die Modellierung naturwissenschaftlicher Kompetenz im Grundschulalter. Projekt Science-P. *Zeitschrift für Pädaogogk, Beiheift*, *56*, 115–125. Retrieved from http://www.pedocs.de/volltexte/2010/3385/

Hartmann, S., Upmeier zu Belzen, A., Krüger, D., & Pant, H. A. (2015). Scientific reasoning in higher education: Constructing and evaluating the criterion-related validity of an assessment of preservice science teachers' competencies. *Zeitschrift für Psychologie / Journal of Psychology*, *223*, 47–53. http://doi.org/10.1027/2151-2604/a000199

Hausmann, R. G. M., van de Sande, B., & VanLehn, K. (2008). Are self-explaining and coached problem solving more effective when done by pairs of students than alone ? In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the cognitive science society* (pp. 2369–2374). Austin: Cognitive Science Society.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*, 77–89. http://doi.org/10.1080/19312450709336664

Hilbert, T. S., Renkl, A., Kessler, S., & Reiss, K. (2008). Learning to prove in geometry: Learning from heuristic examples and how it can be supported. *Learning and Instruction, 18*, 54–65. http://doi.org/10.1016/j.learninstruc.2006.10.008

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist*, *42*, 99–107. http://doi.org/10.1080/00461520701263368

Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Science Education*, *88*, 28–54. http://doi.org/10.1002/sce.10106

Hoogerheide, V., Loyens, S. M. M., & van Gog, T. (2014). Comparing the effects of worked examples and modeling examples on learning. *Computers in Human Behavior*, *41*, 80–91. http://doi.org/10.1016/j.chb.2014.09.013

Hoogerheide, V., Loyens, S. M. M., & van Gog, T. (2016). Learning from video modeling examples: Does gender matter? *Instructional Science*, *44*, 69–86. http://doi.org/10.1007/s11251-015-9360-y

Hoogerheide, V., van Wermeskerken, M., Loyens, S. M. M., & van Gog, T. (2016). Learning from video modeling examples: Content kept equal, adults are more effective models than peers. *Learning and Instruction*, *44*, 22–30. http://doi.org/10.1016/j.learninstruc.2016.02.004

Hoogveld, A. W. M., Paas, F. G. W. C., & Jochems, W. M. G. (2005). Training higher education teachers for instructional design of competency-based education: Product-oriented versus process-oriented worked examples. *Teaching and Teacher Education*, *21*, 287–297. http://doi.org/10.1016/j.tate.2005.01.002

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books.

Jackson, S. L., Stratford, S. J., Krajcik, J. S., & Soloway, E. (1994). Making dynamic modeling accessible to precollege science students. *Interactive Learning Environments*,

4, 233–257. http://doi.org/10.1080/1049482940040305

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, *19*, 509–539. http://doi.org/10.1007/s10648-007-9054-3

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, *38*, 23–31. http://doi.org/10.1207/S15326985EP3801_4

Kalyuga, S., Chandler, P., Tuovinen, J. E., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology*, *93*, 579–588. http://doi.org/10.1037/0022-0663.93.3.579

Kanari, Z., & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching*, *41*, 748–769. http://doi.org/10.1002/tea.20020

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science*, *40*, 651–672. http://doi.org/10.1007/s11251-012-9209-6

Keselman, A. (2003). Supporting inquiry learning by promoting normative understanding of multivariable causality. *Journal of Research in Science Teaching*, *40*, 898–921. http://doi.org/10.1002/tea.10115

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*, 75–86. http://doi.org/10.1207/s15326985ep4102

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*, 1–48. http://doi.org/10.1016/0364-0213(88)90007-9

Klahr, D., Fay, A. L., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology*, *25*, 111–146. http://doi.org/10.1006/cogp.1993.1003

Klahr, D., & Nigam, M. (2005). The equivalence of learninging paths in early science instruction: Effects of direct instrction and discovery learning. *Psychological Science*, *15*, 661–667. http://doi.org/10.1111/j.0956-7976.2004.00737.x

Klahr, D., Triona, L., & Williams, C. (2007). Hands on what? The relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. *Journal of Research in Science Teaching*, *44*, 183–203. http://doi.org/10.1002/tea

Klos, S., Henke, C., Kieren, C., Walpuski, M., & Sumfleth, E. (2008).

Naturwissenschaftliches Experimentieren und chemisches Fachwissen - zwei verschiedene Kompetenzen. *Zeitschrift für Pädagogik, 54*, 304–321.

KMK = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2004). *Bildungsplan 2004: Allgemeinbildendes Gymnasium*. München: Luchterhand.

KMK = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004. Beschlüsse der Kulutministerkonferenz*. München: Luchterhand.

Koenen, J. (2014). *Entwicklung und Evaluation von experimentunterstützten Lösungsbeispielen zur Förderung naturwissenschaftlich- experimenteller Arbeitsweisen. Stuiden zum Chemie- und Physiklernen* (Vol. 171). Berlin: Logos Verlag.

Koerber, S., Sodian, B., Kropf, N., Mayer, D., & Schwippert, K. (2011). Die Entwicklung des wissenschaftlichen Denkens im Grundschulalter: Theorieverständnis, Experimentierstrategien, Dateninterpretation. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 43*, 16–21. http://doi.org/10.1026/0049-8637/a000027

Koerber, S., Sodian, B., Thoermer, C., & Nett, U. (2005). Scientific reasoning in young children: Preschoolers' ability to evaluate covariation evidence. *Swiss Journal of Psychology, 64*, 141–152. http://doi.org/10.1024/1421-0185.64.3.141

Köller, O. (2008). *Evaluation der Standards in den Fächern Biologie, Chemie und Physik für die Sekundarstufe I (ESNaS). Band 5: Aufgabenbeispiele für die Aufgabenentwicklung in den Fächern Biologie, Chemie und Physik für den Kompetenzbereich „Erkenntnisgewinnung"*.

Kontra, C., Lyons, D. J., Fischer, S. M., & Beilock, S. L. (2015). Physical experience enhances science learning. *Psychological Science, 26*, 737–749. http://doi.org/10.1177/0956797615569355

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology. General, 131*, 147–162. http://doi.org/10.1037/0096-3445.131.2.147

Kühl, T., Scheiter, K., Gerjets, P., & Gemballa, S. (2011). Can differences in learning strategies explain the benefits of learning from static and dynamic visualizations? *Computers and Education, 56*, 176–187. http://doi.org/10.1016/j.compedu.2010.08.008

Kuhn, D. (2010). What is scientific thinking and how does it develop? In U. Goswami (Ed.),

Blackwell handbook of childhood cognitive development (2nd ed., pp. 497–523). Oxford: Blackwell Publishing. http://doi.org/10.1002/9780470996652.ch17

Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science*, *16*, 866–70. http://doi.org/10.1111/j.1467-9280.2005.01628.x

Kuhn, D., & Franklin, S. (2006). The second decade: What develops (and how). In R. Damon, R. M. Lerner, D. Kuhn, & R. S. Siegler (Eds.), *Handbook of child psychology: Vol. 2. Cognition, perception and language* (6th ed., pp. 953–993). Hoboken: John Wiley & Sons, Inc. http://doi.org/10.1002/9780470147658.chpsy0222

Kuhn, D., & Pease, M. (2008). What needs to develop in the development of inquiry skills? *Cognition and Instruction*, *26*, 512–559. http://doi.org/10.1080/07370000802391745

Kuhn, D., & Phelps, E. (1982). The development of problem-solving strategies. In H. W. Reese (Ed.), *Advances in child development and behavior: Volume 17* (pp. 1–44). New York: Academic Press.

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, *15*, 11–24. http://doi.org/10.1002/tea.3660150103

Lawson, A. E. (2009). Basic inferences of scientific reasoning, argumentation, and discovery. *Science Education*, *94*, 336–364. http://doi.org/10.1002/sce.20357

Lazonder, A. W., Hagemans, M. G., & de Jong, T. (2010). Offering and discovering domain information in simulation-based inquiry learning. *Learning and Instruction*, *20*, 511–520. http://doi.org/10.1016/j.learninstruc.2009.08.001

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research*, *86*, 681–718. http://doi.org/10.3102/0034654315627366

Leppink, J., Paas, F. G. W. C., van Gog, T., van der Vleuten, C. P. M., & van Merriënboer, J. J. G. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction*, *30*, 32–42. http://doi.org/10.1016/j.learninstruc.2013.12.001

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *29*, 4–16. http://doi.org/10.3102/001318X029002004

Loibl, K., & Rummel, N. (2014). Knowing what you don't know makes failure productive. *Learning and Instruction*, *34*, 74–85. http://doi.org/10.1016/j.learninstruc.2014.08.004

Lowe, R. (2004). Interrogation of a dynamic visualization during learning. *Learning and*

Instruction, 14, 257–274. http://doi.org/10.1016/j.learninstruc.2004.06.003

Mayer, J. (2007). Erkenntnisgewinnung als wissenschaftliches Problemlösen. In D. Krüger & H. Vogt (Eds.), *Theorien in der biologiedidaktischen Forschung: Ein Handbuch für Lehramtsstudenten und Doktoranden* (pp. 177–186). Berlin Heidelberg: Springer. http://doi.org/10.1007/978-3-540-68166-3_16

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *The American Psychologist, 59*, 14–19. http://doi.org/10.1037/0003-066X.59.1.14

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 101*, 343–352. http://doi.org/10.1037/h0043158

Morris, B. J., Croker, S., Masnick, A. M., & Zimmerman, C. (2012). The emergence of scientific reasoning. In H. Kloos, B. J. Morris, & J. L. Amaral (Eds.), *Current Topics in Children's Learning and Cognition* (pp. 61–82). Rijeka: InTech. http://doi.org/10.5772/53885

Mulder, Y. G., Lazonder, A. W., & de Jong, T. (2014). Using heuristic worked examples to promote inquiry-based learning. *Learning and Instruction*, *29*, 56–64. http://doi.org/10.1016/j.learninstruc.2013.08.001

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. (Commitee on the Foundations of Assessment, J. W. Pellegrino, N. Chudowsky, & R. Glaser, Eds.). Washington DC: National Academy Press. http://doi.org/10.17226/10019

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington DC: The National Academies Press. http://doi.org/10.17226/13165

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 26, pp. 125–173). New York: Academic Press. http://doi.org/10.1016/S0079-7421(08)60053-5

Nievelstein, F., van Gog, T., van Dijck, G., & Boshuizen, H. P. A. (2013). The worked example and expertise reversal effect in less structured tasks: Learning to reason about legal cases. *Contemporary Educational Psychology*, *38*, 118–125. http://doi.org/10.1016/j.cedpsych.2012.12.004

Njoo, M., & de Jong, T. (1993). Exploratory learning with a computer simulation for control theory: Learning processes and instructional support. *Journal of Research in Science*

Teaching, *30*, 821–844. http://doi.org/10.1002/tea.3660300803

OECD. (2007). *PISA 2006: Science competencies for tomorrow's world. Volume 1: Analysis*. Paris: OECD. Retrieved from http://www.nbbmuseum.be/doc/seminar2010/nl/bibliografie/opleiding/analysis.pdf

Osborne, J. (2013). The 21st century challenge for science education: Assessing scientific reasoning. *Thinking Skills and Creativity*, *10*, 265–279. http://doi.org/10.1016/j.tsc.2013.07.006

Paas, F. G. W. C., & Sweller, J. (2012). An evolutionary upgrade of cognitive load theory: Using the human motor system and collaboration to support the learning of complex cognitive tasks. *Educational Psychology Review*, *24*, 27–45. http://doi.org/10.1007/s10648-011-9179-2

Paas, F. G. W. C., & van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology*, *86*, 122–133. http://doi.org/10.1037/0022-0663.86.1.122

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (Eds.). (2013). *IQB-Ländervergleich 2012: Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I*. Münster: Waxmann. Retrieved from https://www.iqb.hu-berlin.de/laendervergleich/lv2012/Bericht

Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., … Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, *14*, 47–61. http://doi.org/10.1016/j.edurev.2015.02.003

Pedro, M. S., Gobert, J. D., & Baker, R. (2012). Assessing the learning and transfer of data collection o inquiry skills using educational data mining on students' log files. *Paper Presented at The Annual Meeting of the American Educational Research Association*. Vancouver, BC, CA. Retrieved from http://users.wpi.edu/~rsbaker/SaoPedroetal_AERA2012_FINAL.pdf

Piekny, J., Grube, D., & Maehler, C. (2014). The development of experimentation and evidence evaluation skills at preschool age. *International Journal of Science Education*, *36*, 334–354. http://doi.org/10.1080/09500693.2013.776192

Piekny, J., & Maehler, C. (2013). Scientific reasoning in early and middle childhood: The development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *British Journal of Developmental Psychology*, *31*, 153–179. http://doi.org/10.1111/j.2044-835X.2012.02082.x

Quellmalz, E. S., Davenport, J. L., Timms, M. J., DeBoer, G. E., Jordan, K. A., Huang, C.-W., & Buckley, B. C. (2013). Next-generation environments for assessing and promoting complex science learning. *Journal of Educational Psychology, 105*, 1100–1114. http://doi.org/10.1037/a0032220

Quellmalz, E. S., Timms, M. J., & Schneider, S. A. (2009). *Assessment of student learning in science simulations and games*. Washington DC: National Research Council.

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology, 88*, 144–161. http://doi.org/10.1037//0022-0663.88.1.144

Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *Journal of the Learning Sciences, 13*, 273–304. http://doi.org/10.1207/s15327809jls1303_2

Reisslein, J., Atkinson, R. K., Seeling, P., & Reisslein, M. (2006). Encountering the expertise reversal effect with a computer-based environment on electrical circuit analysis, *16*, 92–103. http://doi.org/10.1016/j.learninstruc.2006.02.008

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cognitive Science, 38*, 1–37. http://doi.org/10.1111/cogs.12086

Renkl, A. (2015). Different roads lead to Rome: The case of principle-based cognitive skills. *Learning: Research and Practice, 1*, 79–90. http://doi.org/10.1080/23735082.2015.994255

Renkl, A., & Atkinson, R. K. (2002). Learning from examples: Fostering self-explanations in computer-based learning environments. *Interactive Learning Environments, 10*, 105–119. http://doi.org/10.1076/ilee.10.2.105.7441

Renkl, A., Atkinson, R. K., Maier, U. H., & Staley, R. (2002). From example study to problem solving: Smooth transitions help learning. *The Journal of Experimental Education, 70*, 293–315. http://doi.org/10.1080/00220970209599510

Renkl, A., Mandl, H., & Gruber, H. (1996). Inert knowledge: Analyses and remedies. *Educational Psychologist, 31*, 115–121. http://doi.org/10.1207/s15326985ep3102_3

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. http://doi.org/10.1006/ceps.1997.0959

Ross, B. H., & Kilbane, M. C. (1997). Effects of principle explanation and superficial similarity on analogical mapping in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 427–440.

http://doi.org/10.1037//0278-7393.23.2.427

Rummel, N., & Spada, H. (2005). Learning to collaborate: An instructional approach to promoting collaborative problem solving in computer-mediated settings. *The Journal of the Learning Sciences*, *14*, 201–241. http://doi.org/10.1207/s15327809jls1402

Rummel, N., Spada, H., & Hauser, S. (2009). Learning to collaborate while being scripted or by observing a model. *International Journal of Computer-Supported Collaborative Learning*, *4*, 69–92. http://doi.org/10.1007/s11412-008-9054-4

Schauble, L., Klopfer, L. E., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching*, *28*, 859–882. http://doi.org/10.1002/tea.3660280910

Scheiter, K., Gerjets, P., & Schuh, J. (2004). The impact of example comparisons on schema acquisition: Do learners really need multiple examples? In Y. B. Kafai, W. A. Sandova, N. Enyedy, A. S. Nixon, & F. Herrera (Eds.), *Proceedings of the 6th international conference on Learning sciences* (pp. 457–464). Mahwah, NJ: Erlbaum. Retrieved from http://dl.acm.org/citation.cfm?id=1149182

Schmidt-Weigand, F., & Scheiter, K. (2011). The role of spatial descriptions in learning from multimedia. *Computers in Human Behavior*, *27*, 22–28. http://doi.org/10.1016/j.chb.2010.05.007

Schunk, D. H. (1987). Peer models and children's behavioral change. *Review of Educational Research*, *57*, 149–174. http://doi.org/10.3102/00346543057002149

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, *22*, 129–184. http://doi.org/10.1207/s1532690xci2202_1

Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review*, *39*, 37–63. http://doi.org/10.1016/j.dr.2015.12.001

Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Aleven, V., & Salden, R. (2009). The worked-example effect: Not an artefact of lousy control conditions. *Computers in Human Behavior*, *25*, 258–266. http://doi.org/10.1016/j.chb.2008.12.011

Schworm, S., & Renkl, A. (2007). Learning argumentation skills through the use of prompts for self-explaining examples. *Journal of Educational Psychology*, *99*, 285–296. http://doi.org/10.1037/0022-0663.99.2.285

Seidel, T., Blomberg, G., & Renkl, A. (2013). Instructional strategies for using video in teacher education. *Teaching and Teacher Education*, *34*, 56–65.

http://doi.org/10.1016/j.tate.2013.03.004

Shahali, E. H. M., & Halim, L. (2010). Development and validation of a test of integrated science process skills. *Procedia - Social and Behavioral Sciences*, *9*, 142–146. http://doi.org/10.1016/j.sbspro.2010.12.127

Siegler, R. S. (1989). Mechanisms of cognitive development. *Annual Review of Psychology*, *40*, 353–79. http://doi.org/10.1146/annurev.ps.40.020189.002033

Siegler, R. S., & Chen, Z. (1998). Developmental differences in rule learning: A microgenetic analysis. *Cognitive Psychology*, *36*, 273–310. http://doi.org/10.1006/cogp.1998.0686

Siegler, R. S., & Liebert, R. M. (1975). Acquisition of formal scientific reasoning by 10- and 13-year-olds: Designing a factorial experiment. *Developmental Psychology*, *11*, 401–402. http://doi.org/10.1037/h0076579

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development*, *62*, 753–766.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*, 257–285. http://doi.org/10.1016/0364-0213(88)90023-7

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, *2*, 59–89. http://doi.org/10.1207/s1532690xci0201_3

Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*, 251–296. http://doi.org/10.1023/A:1022193728205

Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology*, *28*, 129–160. http://doi.org/10.1016/S0361-476X(02)00011-5

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, *95*, 66–73. http://doi.org/10.1037/0022-0663.95.1.66

Tomlinson, P. D., & Hunt, D. E. (1971). Differential effects of rule-example order as a function of learner conceptual level. *Canadian Journal of Behavioural Science*, *3*, 237–245. http://doi.org/10.1037/h0082265

Toth, E. E., Suthers, D. D., & Lesgold, A. M. (2002). "Mapping to know": The effects of representational guidance and reflective assessment on scientific inquiry. *Science Education*, *86*, 264–286. http://doi.org/10.1002/sce.10004

Trafton, J. G., & Reiser, B. J. (1993). The contribution of studying examples and solving

problems to skill acquisition. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society* (pp. 1017–1022). Hillsdale, NJ: Lawrence Erlbaum Associates.

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. San Francisco: Jossey-Bass.

Triona, L., & Klahr, D. (2003). Point and click or grab and heft: Comparing the influence of physical and virtual instructional materials on elementary school students' ability to design experiments. *Cognition and Instruction, 21*, 149–173. Retrieved from http://www.tandfonline.com/doi/abs/10.1207/S1532690XCI2102_02

Trundle, K. C., & Bell, R. L. (2010). The use of a computer simulation to promote conceptual change: A quasi-experimental study. *Computers and Education, 54*, 1078–1088. http://doi.org/10.1016/j.compedu.2009.10.012

Tuovinen, J. E., & Sweller, J. (1999). A comparison of cognitive load associated with discovery learning and worked examples. *Journal of Educational Psychology, 91*, 334–341. http://doi.org/10.1037/0022-0663.91.2.334

van der Meij, H., & van der Meij, J. (2013). Eight guidelines for the design of instructional videos for software training. *Technical Communication, 60*, 205–228. Retrieved from http://www.ingentaconnect.com/content/stc/tc/2013/00000060/00000003/art00004

van Gog, T., Kester, L., & Paas, F. G. W. C. (2011). Effects of worked examples, example-problem, and problem-example pairs on novices' learning. *Contemporary Educational Psychology, 36*, 212–218. http://doi.org/10.1016/j.cedpsych.2010.10.004

van Gog, T., & Rummel, N. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review, 22*, 155–174. http://doi.org/10.1007/s10648-010-9134-7

van Joolingen, W., & de Jong, T. (1993). Exploring a domain with a computer simulation: Traversing variable and relation space with the help of a hypothesis scratchpad. In D. Towne, T. de Jong, & H. Spada (Eds.), *Simulation-Based Experiential Learning* (pp. 191–206). Berlin, Germany: Springer-Verlag. http://doi.org/10.1007/978-3-642-78539-9_14

van Loon, M. H., de Bruin, A. B. H., van Gog, T., van Merriënboer, J. J. G., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta Psychologica, 151*, 143–154. http://doi.org/10.1016/j.actpsy.2014.06.007

van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning:

Recent developments and future directions. *Educational Psychology Review*, *17*, 147–177. http://doi.org/10.1007/s10648-005-3951-0

Wecker, C., Rachel, A., Heran-Dörr, E., Waltner, C., Wiesner, H., & Fischer, F. (2013). Presenting theoretical ideas prior to inquiry activities fosters theory-level knowledge. *Journal of Research in Science Teaching*, *50*, 1180–1206. http://doi.org/10.1002/tea.21106

Wellnitz, N., Fischer, H. E., Kauertz, A., Neumann, I., & Pant, H. A. (2012). Evaluation der Bildungsstandards – eine fächerübergreifende Testkonzeption für den Kompetenzbereich Erkenntnisgewinnung. *Zeitschrift für Didaktik der Naturwissenschaften*, *18*, 261–292.

Wilkening, F., & Sodian, B. (2005). Scientific reasoning in young children: Introduction. *Swiss Journal of Psychology*, *64*, 137–139. http://doi.org/10.1024/1421-0185.64.3.137

Zacharia, Z. C., Olympiou, G., & Papaevripidou, M. (2008). Effects of experimenting with physical and virtual manipulatives on students' conceptual understanding in heat and temperature. *Journal of Research in Science Teaching*, *45*, 1021–1035. http://doi.org/10.1002/tea.20260

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, *41*, 64–70. http://doi.org/10.1207/s15430421tip4102_2

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review*, *20*, 99–149. http://doi.org/10.1006/drev.1999.0497

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*, 172–223. http://doi.org/10.1016/j.dr.2006.12.001