

SUPPORTING EVIDENCE EVALUATION SKILLS

Process support for evaluating sinking objects in an authentic inquiry learning environment

Master thesis Educational Science and Technology

Carolin Richtering
University of Twente
May 2012

Supervised by
Dr. Ard W. Lazonder
Dr. Pascal Wilhelm

Abstract

Inquiry learning can promote the acquisition of science content if students are adequately supported during their investigations. The evaluation of evidence lies at the heart of the inquiry process. Given the many difficulties young students experience with observing, interpreting, and generalizing from evidence, these skills require additional support. The present study examined whether process support prompts can enhance students' performance of evidence evaluation skills and hence the acquisition of science content. Fifty-seven seventh graders from a comprehensive school took part in an authentic inquiry task about sinking objects where misconceptions were likely to exist. Students in both experimental conditions received process support prompts that structured, exemplified, and problematized the initial task. For students in the full support condition, this support addressed the skills involved in observation, interpretation, and generalization. Students in the observation support condition only received the prompts that supported observation skills, whereas students in the control condition received no support at all. Overall results were disappointing. The mastery of evidence evaluation skills could only be promoted to some extent by the support. Gain in science content could not be detected. These findings raise questions about more adequate ways of implementing process support and the conditions under which better evidence evaluation skills lead to a gain in science content.

Teaching science is not easy. Scientific concepts are hard to impart and even after years of science education, people still retain naïve conceptions about scientific phenomena. Children in particular hold rather persistent misconceptions about scientific phenomena (Chinn & Brewer, 1993), which are generally difficult to correct by traditional methods of science teaching (Sodian, 1998). Learning science content therefore involves more than extending existing knowledge schemata: it requires a change in core concepts of the theory and their interrelation (Vosniadou & Brewer, 1987). This kind of knowledge restructuring is known as conceptual change. Chinn and Malhotra (2002a) argue that significant conceptual change is only realized when learners also generalize their findings to other situations and a long-lasting change in conception takes place, otherwise prior beliefs and theory might reemerge after some time.

In a recent meta-analysis, Alfieri, Brooks, Aldrich, and Tenenbaum (2011) concluded that inquiry learning is more beneficial than explicit instruction, provided that learners are supported during their investigations. In inquiry learning, students actively construct knowledge by engaging in activities such as generating hypotheses about scientific phenomena, designing and executing experiments, and evaluating evidence. Inquiry learning is based on Dewey's beliefs that science education should not be so much about "ready-made knowledge" or "subject-matter of fact and law" (Dewey, 1910/1995, p. 394), but should replicate the experience of science. For the promotion of scientific understanding, evidence evaluation is the central activity during inquiry learning (Kuhn, 1989; Zimmerman, 2007). Evidence evaluation is composed of three consecutive sub-skills: observation, interpretation, and generalization. Learners observe the outcomes of experiments closely before drawing inferences and generalizing their findings to similar situations (Chinn & Malhotra, 2002a).

However, inquiry learning in general, and evidence evaluation in particular, are challenging to young learners. Making correct observations is often problematic because learners' perception is unsystematic and unfocused and "across settings and ages children seem predisposed to arbitrarily noticing phenomena" (Eberbach & Crowley, 2009, p. 48). Many inquiry tasks also involve the observation of very small differences in physical quantities such as time or weight. In an inquiry task about buoyant forces, Penner and Klahr (1996) had learners compare the sinking times of various objects. Minimal differences in weight (e.g., 0.9 and 0.5 grams) and sinking time (e.g., 0.83 and 0.92 seconds) had to be observed. Learners experienced the same difficulties in observing evidence in the rock-dropping task (Chinn & Malhotra, 2002a), the canal task, and the spring task (Schauble, 1996). The classification of observations can also be problematic for learners. Eberbach and Crowley found in their study (2009) that many children attend to surface features, familiar objects, perceive instances in isolation, and get lost in unimportant details because they still lack the content knowledge that would allow them to access deeper features and control their attention. However, when learners hold correct prior beliefs, they are more likely to make correct judgments (Chinn & Malhotra, 2002a; Richter, 2010). A possible explanation is that with correct predictions, they focus their attention accordingly which increases the chance of correctly observing presented stimuli. Children furthermore do not spontaneously take notes (Eberbach & Crowley, 2009; Schauble, 1990), a means of streamlining observed results by omitting irrelevant information and highlighting theoretically important features.

During interpretation, data obtained from observations and prior knowledge are ideally assessed interdependently by learners (Chinn & Brewer, 2001). Interpreting data can be problematic because, although learners make many observations, they have difficulty encoding, making valid inferences and connecting observations to theory. Children lack the skills to extract relevant data features from noise and error. This is why Eberbach and Crowley (2009) call children "dust-bowl empiricists". Children also tend to make judgments based on inconclusive or insufficient evidence. Right from the start of their inquiry, they focus on making deterministic causal inferences, regardless of whether these are warranted or not (Zimmerman, 2007). More severe problems arise when the relationship between variables is more complex. Children have great problems detecting interaction effects between independent variables (Schauble, 1996). Although more than half of the participants in Schauble's study conducted experiments that were relevant to discovering interactions between variables, most children did not detect these effects. Beishuizen, Wilhelm, and Schimmel (2004) report that a full understanding of interaction does not arise normally before the age of sixteen. When learners assess the validity of data, prior beliefs have to be bracketed. When children have rich

theories and hold strong prior beliefs, they have difficulties in setting these beliefs aside (Koslowski, 1996). When observational data is anomalous and evidence conflicts with incorrect prior beliefs, learners have to change their pre-existing beliefs. Chinn and Brewer (1998) and Lin (2007) empirically tested how undergraduate students react to anomalous data. Only a rather small percentage of about 10% of the learners actually changed their theories; the remaining learners rejected the anomalous data for a variety reasons.

Generalizing what has been observed and interpreted to situations that are dissimilar in some respects of the experimental situation, also referred to as near transfer, does not always take place without problems. Chinn and Malhotra (2002a) found in studies with eight to twelve year olds that generalizing inferences drawn from observational data to other situations went smoothly when prior beliefs were confirmed. However, when observations contradicted prior beliefs, only two third of the inferences were generalized.

Together these studies show that learners' difficulties with the sub-skills of evidence evaluation are diverse. There are problems with perception, classification, attention and note taking during observation; problems with encoding, making valid inferences and connecting observations to theory, especially when variable relationships are complex or a change in conception has to take place; and problems with transferring inferences when prior beliefs are contradicted during generalization.

Adequate support for evidence evaluation skills

Despite the challenging nature of inquiry tasks and in particular the performance of evidence evaluation skills, learners can benefit from inquiry learning when they receive adequate support (Kirschner, Sweller & Clark, 2006; Alifieri et al., 2011). A division in content support, scaffolding learners by providing new or activating prior knowledge, and process support, scaffolding learners by improving inquiry skill performance, has proven as a helpful classification for inquiry support. Chinn and Malhotra (2002a) examined these two ways of supporting learners during a rock-dropping task. They provided content support by giving learners a scientific explanation before letting them evaluate evidence. They also supported the process of evidence evaluation with data-based discussions with pairs of learners before the inquiry activity that let learners reflect on the interpretation of data. Chinn and Malhotra found content support to be effective in terms of learning outcomes. The data-based discussion, however, did not show any effect. Other studies have confirmed that activating existing beliefs and theories (Fund, 2007) or providing learners with relevant content knowledge (Lazonder, Wilhelm & Hagemans, 2008) results in better learning.

But there are, despite the effectiveness of content support, many arguments against employing this kind of support. Supporting learners with content support is not very supportive in the long-run because it leads to a schema-driven way of investigation and does not promote the development of inquiry skills (Chinn & Malhotra, 2002a). Other researchers also found that even learners with rather extensive knowledge of the domain were not capable of performing scientific reasoning skills proficiently (e.g., Mulder, Lazonder & de Jong, 2010). A more principled objection is that content support conflicts with the inherent concept of inquiry learning: learners should discover content knowledge themselves (Lazonder, Hagemans & de Jong, 2010). Therefore, the idea of process support, enabling learners to perform activities they would not be able to do without the support and in the long run and with repeated practice resulting in improved inquiry skills, should be reconsidered. A different design of process support could possibly enhance its effectiveness. The implementation of process support can be potentially promoted by giving it just in time instead of in advance of the activity. Support could furthermore take a broader approach and support all skills of evidence evaluation where support is needed, not only at a limited number as with the data-based discussion of Chinn and Malhotra (2002a). Process support could also be more attuned to the kind of difficulties learners experience.

The divergent difficulties learners have with the sub-skills of evidence evaluation described above lead one to conclude that a thorough support of all sub-skills is necessary. This might however not be the case. Chinn and Malhotra (2002a) could locate observations skills as the bottleneck of evidence evaluation, meaning that problems during subsequent stages of evidence evaluation are mainly caused by difficulties during observation. According to this logic, it is redundant for adequate evidence evaluation skills support to addresses all three sub-skills of evidence evaluation. It would be sufficient to support observation skills only; baldly said "when observation goes well the rest will

too". An additional reason to merely support observation skills is that too much support might impede the learning process. This might be the case when learners have no experience at all with evaluating evidence and are overwhelmed with the support. But even when their experience increases, process support might lose its effectiveness or, worse, have a negative effect on learning. This phenomenon is known as the expertise-reversal effect (Kalyuga, Ayres, Chandler & Sweller, 2003). Learners with little experience in evidence evaluation benefit from process support strategies while it is redundant or even disruptive for more experienced learners. Redundant support is hard to ignore and requires additional mental effort.

Process support that is attuned to the kind of difficulties learners face, addresses each difficulty with an adequate strategy. A process support strategy that helps learners with the complexity of evidence evaluation skills structures the inquiry activity by taking over routine aspects of the task, or aspects that are still too difficult in order to let learners devote energy to more important aspects that are appropriate for their skill level (Reiser, 2004). The complexity of the task can be reduced by narrowing down options, preselecting data, or decomposing the task. Fund (2007) found, for example, that seventh graders who were supported by a worksheet prompting them to direct attention to important data showed better learning outcomes. A second process support strategy is to make sure that learners are provided with extensive practice for skills that are not routine or too difficult but fundamental for evidence evaluation. This strategy works by exemplification and is a very straightforward support strategy. Learners "should be explicitly shown what to do and how to do it" (Kirschner et al., 2006, p.79). Beishuizen and colleagues (2004), for example, found that sixth graders in a training group were more successful in detecting interaction effects than pupils in a practice-only group. A third process support strategy is problematizing aspects of the task so as to stimulate learners to attend to issues they might otherwise not (consciously) address (Reiser, 2004). Such issues can be as diverse as ambiguous experiments, complex causality, experiments with invalid or restrictedly valid data, or experiments where bias is likely. Problematising either works by marking critical features, by inducing cognitive conflict, or by emphasizing discrepancies between what the learners' performance and correct performance. It is therefore not surprising that problematising aspects of the task initially makes the accomplishment of the task more difficult. However, confronting learners with problematic situations is productive for learning in the long run (Reiser, 2004). To address the variety of challenges learners face when evaluating evidence, a comprehensive approach that combines the three process support strategies (structuring, exemplifying and problematising) ensures that children get the right amount of scaffolding. The effectiveness of this approach will manifest itself in better performance during evidence evaluation, and in the long run, with repeated practice, in improved evidence evaluation skills.

Adequate support for evidence evaluation skills does not only promote skill performance. It has often been argued that inquiry skills are pivotal to the construction of knowledge schemata. This is a logical implication of the bootstrapping idea (e.g., Koslowski, 1996; Lehrer, Schauble & Lucas, 2008), which states that content knowledge and inquiry skills are not developing independently; mastery of one bootstraps the other. A more sophisticated set of inquiry skills thus promotes the generation of a deeper, more organized knowledge base.

Present study

In order to support evidence evaluation skills in an adequate manner, measures that aim at supporting the inquiry process are reconsidered. The present study departs at the difficulties young learners have with the sub-skills of evidence evaluation. The study employs process support that consists of a combination of structuring, exemplifying and problematising strategies. It was examined to what extent this type of process support can enhance the performance of evidence evaluation skills and thereby the acquisition of science content.

However, providing process support for all sub-skills of evidence evaluation might be redundant. According to Chinn and Malhotra (2002a), who define observation skills as the bottleneck of evidence evaluation, supporting these skills only is sufficient. More support is likely to overstrain learners. A second research question of this study was therefore whether it is more effective to support observations skills only.

Expectations for the present study were that process support will enhance skill performance and by that the learning of deep and meaningful science knowledge, compared to a control group

without any support. Supporting only observations skills should, while being less time consuming, lead to even better learning results.

Method

Participants

Fifty-seven seventh grader from a German comprehensive school participated in the study. Participants were randomly assigned to one of the three experimental conditions. Nine participants had to be removed from the sample for various reasons. One child turned off the computer and another one the log program, despite repeated and insistent instructions not to do so. One child had to abandon the study due to family reasons, and another one did not succeed to finish. One of the two classes was tested in the week before the start of the summer holidays. One child in this class refused to participate, and four other children did not take the test seriously. The final sample therefore consisted of 48 participants: 15 in the full support condition, 17 in the observation support condition, and 16 in the no support condition.

Inquiry task

Research into inquiry learning increasingly employs task with features of authentic science (Chinn and Malhotra, 2002b). Authentic task are especially important for practicing inquiry skills because task too remote from real practice only address oversimplified skills that have little in common with real-world scientific reasoning. Authentic inquiry task are based on a complex underlying model of variables which often interact and do not follow a linear relationship. Such tasks reflect the ‘messiness’ of the natural world by containing ambiguous data and experimental flaws. Following this line of reasoning, participants in this study worked on an inquiry task about sinking objects, a topic from their physics curriculum. This task was chosen because it required learners to not only learn something new, but also to reconstruct existing knowledge schemata. A common misconception, even among adults, is that the weight of an object is the most single, important factor for predicting sinking time (Penner & Klahr, 1996).

The inquiry task was presented in a computerized learning environment. In this environment, learners evaluated evidence by watching a series of 30 video-recorded experiments. In the experiments, solid objects of different materials, form and size (see Table 1) were dropped in 90 cm cylinders filled to 85 cm with tap water. The objects were painted (stainless steel in red, glass in blue) in order to create similar surface textures and to make them easier to discern. Learners had to determine which of the two dropped objects reached the bottom of the cylinder first. As learners could not weigh and measure the video-recorded objects, they were given an overview of the properties of each object prior to the start of an experiment and additionally in a section of the learning environment.

Table 1

Dimension, weight, density, sink time and order of sinking of each object

Material and other attributes ^a	Sphere		Cube		Cuboid	
	Small ø=1 .52 cm ³	Large ø=1,5 1.77 cm ³	Small a=0.8 .51 cm ³	Large a=1.2 1.74 cm ³	Small 1.2x0.8x0.5 0.48 cm ³	Large 2.1x1.2x0.7 1.76 cm ³
Stainless steel						
Weight (g)	4.189	14.137	4.096	13.824	3.84	14.112
Density (g/cm ³)	8					
Sink time (sec)	.69	.60	.92	.79	1.03	.88
SD	.02	.04	.04	.04	.03	.04
Order of sinking	2 nd	1 st	5 th	3 rd	6 th	4 th
Glass						
Weight (g)	1.309	4.418	1.28	4.32	1.2	4.41
Density (g/cm ³)	2.5					
Sink time (sec)	1.41	1.17	1.92	1.56	2.18	1.96
SD	.09	.12	.07	.06	.05	.05
Order of sinking	8 th	7 th	10 th ^b	9 th	12 th	11 th ^b

^a Measured in centimeter.^b Due to variations in sink times, it could not be definitely determined whether the small glass cube or the large glass cuboid sinks faster.

The underlying model of the inquiry task was rather complex. As an object sinks, its acceleration is determined by the difference between the downward force of gravity and the countervailing forces of friction and buoyancy. The velocity at which the object sinks increases until the gravitational force equals the combined forces of friction and buoyancy. The object then sinks at a constant, terminal velocity. Buoyant forces can be calculated by the following formula $F_B = \rho_f \cdot V_{disp} \cdot g$, where ρ_f is the density of the fluid, V_{disp} is the volume of the displaced body of liquid, and g is the gravitational acceleration at the location in question, which can be approximated by 9.81 m/s². When the object is completely submerged under water, the weight of the liquid an object displaces is equal to the force that is pushing it up. It can therefore be stated that the sinking time of an object depends on the difference between the density of the fluid and the density of the object. The frictional force of an object in water, however, cannot be precisely calculated because it is determined by several additional factors. These include the object size, its winding area, its surface texture, and the viscosity of the fluid. In addition, the frictional force increases when the velocity of an object increases.

In total, there were twelve different objects, a sphere, a cube and a cuboid from stainless steel and glass, each in a small and a large version. Sink times of each object were approximated by letting each object sink ten times (see Table 1). A stepwise multiple regression analysis was performed to determine the influence different attributes have on objects' sink time. Material alone attributed for 73% of the variance in sink times, reflecting the relative impact of density. The adjusted R² values yielded that all design attributes should be included in the model, explaining together 96% of the variance. The regression equation for predicting the sink time of an object can be established by a constant of .43 ms, plus (where applicable) adding .88 ms for glass, .33 ms for cube, .54 ms for cuboid and .20 ms for small. Weight alone explained only 38% of the variance.

Counting all unique combinations of the twelve objects, the experiment space comprised 78 experiments. It was ensured that learners would encounter enough meaningful experiments that are essential for learning. From the 30 experiments in the inquiry activity, six experiments, where only one attribute was varied (Control-of-Variables Strategy; CVS) were chosen for each material (e.g. a small stainless steel sphere and a small glass sphere), form (e.g. a large steel cube and a large steel cuboid), and size (e.g. a small glass cube and a large glass cube). Furthermore, three experiments were chosen that violated the misconception that heavier objects always sink faster. These three were especially striking because both objects had the same form (e.g., a small stainless steel cuboid and a large glass cuboid). The remaining experiments were chosen at random. It turned out that in all of them more than one attribute varied. Two experiments also violated the misconception that heavier objects always sink faster. In designing the learning environment, all selected experiment were conducted three times and videotaped. All 90 videos were included in the learning environment to give learners the chance to see an experiment being repeated. One third of the experiments contained slight

flaws due to authenticity. A diverging start time because the experimenter did not always succeed in dropping objects at exactly the same moment is an example of such a flaw, or an ambiguous situation that arises because objects bounce when they hit the bottom or do not sink in a straight trajectory but drift and slither along the wall of the cylinders or sink in a wiggly way.

Learning environment

The learning environment was programmed with the Adobe® Flash® Professional CS4 software. It had the same basic functions for all experimental conditions. Learners were guided consecutively through all 30 experiments. They were first shown the attributes of the two objects under investigation. Learners then started the first video by clicking on the button labeled “start film” and observed the experiment. They then had the option to watch the second and third video. The learning environment had a section where the twelve objects were displayed at all time, see Figure 1. Hovering with the mouse over the objects let a small window with the attributes of the objects pop up. The learning environment also contained a notepad for learners to record the outcomes of the experiments. These notes were then saved under the current experiment and could be revisited at all times. The three versions of the learning environment only differed with respect to the prompts that were used to give process support, which will be described in the next paragraph.

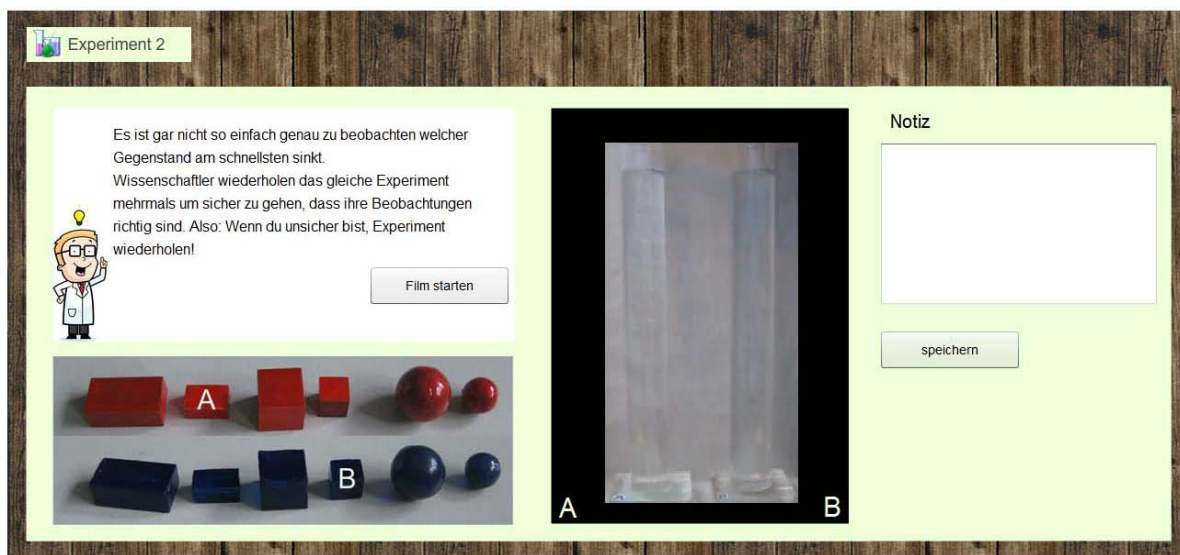


Figure 1. Screenshot of the learning environment, showing an exemplification prompt (top-left), the video section (middle), and the note taking section (right).

Process support

In the full support and in the observation support condition, process support was implemented in the learning environment by means of prompts: short messages suggesting learners what to do. It has been demonstrated that even very simple and repeated prompts can lead to an improvement in inquiry skills (Zimmerman, 2007). A great advantage over data-based discussion or other ways of supporting the process of evidence evaluation is that prompts can be given temporally close to the performance of a skill. Prompts supporting interpretation and generalization skills were presented during the execution of the skill; prompts supporting observation skills were presented shortly before or after skill performance, because one cannot read prompts and observe at the same time. All process prompts were embedded in the learning environment through pop-ups and text fields (see Figure 1 for an example). Learners were focused on the prompts, for example by delaying the next action by some seconds so that they were required to read the prompt. Which prompts were presented to the learners depended on the condition they were in. Learners in the full support condition received prompts that scaffolded all evidence evaluation sub-skills, learners in the observation support condition only received the prompts that addressed observation skills, and learners in the no support condition received no prompts at all.

To address all difficulties learners have with the observation, interpretation, and generalization of evidence (and to explicitly take into account the nature of these difficulties), three different types of prompts were used for each of the evidence evaluation sub-skills. These were structuring prompts, exemplification prompts, and problematizing prompts. Structuring prompts were most prominent during the inquiry activity due to the fact that many routine aspects waste valuable energy and some aspects of task had to be taken over because they are too difficult for learners. This was achieved by reducing the complexity of the task by narrowing down options, decreasing choices, and decomposing the task. The exemplary prompts 1 and 2 shown in Table 2 were present at every observation in both experimental conditions. Prompts 3 and 4 were present at every interpretation and generalization in the full support condition. There were two additional structuring prompts which strengthened prompts 3 and 4.

Table 2

Examples of structuring, exemplification or problematizing prompts

Type of prompt	Example(s)
Structuring	<p>(1) Which object will sink faster? (<i>Before every observation</i>)</p> <p>a) Object A b) Object B c) Both objects will sink at the same rate d) No idea</p> <p>(2) Which object sank faster? (<i>After every observation</i>)</p> <p>a) Object A b) Object B c) Both objects sank at the same rate d) I'm not sure and want to repeat the experiment</p> <p>(3) Why did object A sink faster?/ Why did object B sink faster?/ Why did both objects sink at the same rate? [textInput] (<i>At every interpretation</i>)</p> <p>(4) Based on your observations from the previous experiments, which attributes influence the speed of an object sinking in water? [textInput] (<i>After every fifth experiment, at every generalization</i>)</p>
Exemplification	The correct explanation for an observation is often only discovered when the matter is approached in an objective way. So, try to look at your observation as if you had no prior knowledge and don't exclude any explanation from the outset because it seems very far off. (<i>During interpretation</i>)
Problematizing	When the sinking time of two objects differs only minimally, it is particularly difficult to observe which object sinks fastest. That is completely normal. The human ear or the human eye is not able to perceive such small differences. This does however not mean that no difference exists. (<i>After observation</i>)

Exemplification prompts were deliberately spread over all experiments to offer sufficient opportunity for training aspects of skills that were not routine or too difficult but fundamental for evidence evaluation. Exemplification prompts, like the example shown in Table 2, simply prompted learners what to do and how to do it. There were nine additional exemplification prompts in the full support condition, of which only the four observation prompts were included in the observation support condition.

Problematizing prompts were shown in special situations. This was the case when differences in sinking times were minimal (see Table 2), when results were surprising and therefore bias was likely, when flaws due to authenticity occurred, and in general because of the complex causality underlying the inquiry task. Problematizing prompts supported learners by pointing to otherwise unattended issues and by emphasizing discrepancies between what learners do and correct performance. There were five additional problematizing prompts in the full support condition of which one observation prompts was also included in the observation support condition.

Instruments

The primary goal of process support was to enhance the performance of evidence evaluation skills, to help learners to do something they would not be able to do without the support. However, measuring the mastery of evidence evaluation skills is difficult because most processes take place in

mind. Whether learners observe, interpret or generalize correctly can only be deduced from indirect measures. Process support in the observation support and full support condition used prompts. So, any form of asking question to measure the mastery of evidence evaluation skills would interact with the exact process that is under study (Wilhelm & Beishuizen, 2004). Thus only non-intrusive techniques could be used to measure the mastery of evidence evaluation skills. The present study used log files that recorded every action learners made in the inquiry learning environment and every text input that was made after a prompt to get a general impression of learners' proficiency in evidence evaluation skills. This data was further augmented by notes that were taken spontaneously, thus without prompting.

Learners' knowledge about sinking objects was assessed by a science content pre- and posttest. The pretest contained seven knowledge items, three far generalization items, and three abstract generalization items. The knowledge items addressed task-specific knowledge about sinking objects. The test started with the open question "Which factors determine how fast an object sinks in water?" to examine learners' general idea of sinking objects. On the next page, learners had to arrange all 12 objects from the learning environment (see Table 1) according to their sinking time. This question was followed by five pairwise comparisons to assess what learners think about the influence of form, size, and type of material and, hence, weight. Each pairwise comparison contained a closed question asking learners which of the objects would reach the bottom first, and a subsequent open question asking them to motivate their answer. In the first three comparisons only one attribute varied; in the fourth pairwise comparison size and material varied and weight did not determine the correct answer, and in the fifth comparison all attributes varied. The generalization items aimed to assess whether learners' knowledge of sinking objects is deep-rooted and well organized. Three far generalization items measured generalization of knowledge beyond the scope of the inquiry task. These items consisted of pairwise comparisons of objects with materials, forms and sizes that differed slightly from the ones in the inquiry task. Three additional items assessed generalization on an abstract rule level. The first item was an open question that dealt with the density of peeled and unpeeled oranges. The other two items contained pairwise comparison between objects that were very different from the ones in the inquiry task, with different form, size, and material (but the same weight). Both comparisons were followed by an open question asking learners to explain their answer. A sample question of each item category can be found in Appendix A. The post-test contained the same items in reverse order.

Procedure

A pilot test with two children from the same age group was conducted to examine the understandability of the science content test and the prompts. Test items and prompts that lacked clarity were adapted.

In total, the experiment took half a day. Participants in the study came from two classes. Upon arriving in the computer room, they were told that they were about to test one of three different learning environments about sinking objects that used a new way of learning. Further, they were told that their knowledge about sinking objects would be anonymously tested before and afterwards to see how good the learning environments were. After an introduction to the learning environment, participants completed the science content pretest, which took about twelve minutes. They were then randomly assigned to one of the three experimental conditions. Each pupil was seated in front of one computer, equipped with a headphone. Participants from the same conditions were grouped together. The learning environment guided each participant through a predefined sequence of the 30 selected experiments. Participants in the process support condition received a version of the environment that contained all prompts, participants in the observation support condition worked with a version that only contained the observation prompts, whereas participants from the no support condition performed the inquiry activity without any prompts. All three conditions had the same time to perform experiments and evaluate evidence. When a participant had finished the inquiry activity, s/he was directly presented with the science content post-test. A short debriefing session was held after all learners were finished.

Coding and scoring

Main variables under investigation in this study were mastery of evidence evaluation skills and gain in science content. The level of evidence evaluation skill performance was determined by the number of evidence evaluation skills performed, off-task remarks, the quality of evidence evaluation skill performance, time on task, and the number of repeated experiment.

Logged user input and notes were categorized as performance of evidence evaluation skills (predictions, observations, remarks on flaws due to authenticity, interpretations, and generalizations) or as off-task remarks and counted. In order to be able to compare results, these numbers were averaged per participant. Table 3 shows according to which rules the content of logged user input and notes was categorized.

Table 3
Categorization of logged user input and notes

Categories	Content of logged user input and notes
Evidence evaluation skills	
Observation	
Prediction	Statement expressing prior beliefs about the outcome of one experiment
Observation	Statement describing the outcome of an experiment
Remarks on flaws due to authenticity	Statement about sinking times only differing minimally, objects that were not dropped at the same time, bounce, drift, or slither
Interpretation	Statement that make sense of the outcomes of one experiment
Generalization	Statement about the influence of attributes in general (“size does always...”) and not about one particular experiment, possibly referring to the results of several previous experiments
Off-task remarks	Utterance that has nothing to do with the experiment(s), e.g. “hello”

Insights into the quality of conducted skills were gained by judging the correctness of predictions and observations. The quality of interpretations was estimated per experiment by allocating one point per correct attribute (material, form, size, and weight). Depending on the experiment up to four points were possible which adds up to 62 points in total. Generalization notes were considered in total, and one point was given per correctly mentioned attribute, leading up to a possible total of four points.

Log files from all three conditions provided information on how much time participants needed to finish the inquiry activity, shedding light on the question whether supporting only observation skills is less time consuming. Log files also revealed how many experiments were repeatedly watched. Repeating some experiments (those which are difficult to observe) could be seen as appropriate behavior in opposition with not repeating any experiment or repeating the majority of all experiments.

Gain in science content was determined by the pretest and posttest scores on the knowledge, far and abstract generalization items in the science content test. Participants’ answers on the science content test were checked against the model underlying the inquiry task. The closed questions of the science content test were scored as true or false. For the open questions, a coding scheme was developed based on the underlying physical model. One point was allocated for each correctly mentioned attribute (material, form, size, and weight). For the open questions subsequent to the pairwise comparisons, points were only given when the associated closed question was answered correctly. Two raters coded the open questions from the pre- and posttest of five randomly selected participants from each experimental condition. The inter-rater agreement was 0.87 (Cohen’s *k*).

A special case was the scoring of question 2. The arrangement of objects according to their sinking time was compared with the computed order from the stepwise regression analysis. The more a participant sequenced the objects correctly and the closer s/he placed them to their correct neighbors (tested with both arithmetical correct orders; the higher score counts), the more points s/he received. For example, when a participant sequenced the objects 1 8 2 4 9 7 6 11 5 3 12 10, the discrepancy with the correct order is 0 6 1 0 4 1 1 3 4 7 1 2 (adding up to 30) and the distance to the correct neighbors 0 6 6 1 4 2 1 4 6 2 8 2 (adding up to 42). The two numbers were averaged ($(30+42)/2=36$), subtracted from the worst possible score ($72.5-36=36.5$) and translated into points ($=3.65$).

Design en data analyses

The first set of analyses focused on the mastery of evidence evaluation skills. Data obtained from participant's notes could not be compared by inferential statistical analysis because of the incomparable circumstances: when not prompted, participants were not required to write down predictions, observations, remarks on flaws due to authenticity, interpretations, or generalizations. Still, participants might have performed these skills in mind without taking notes. The number of performed evidence evaluation skills (vs. off-task remarks) and their quality were therefore discussed in qualitative terms. For both experimental conditions a one-way ANOVA could be conducted on the correctness of prediction and observation scores. Participants in these two conditions had to do a prediction and observation because these were prompted. Differences between conditions concerning time on task and number of repeated experiments were also tested with a one-way ANOVA.

A second set of analyses assessed gain in science content. Toward this end a 2 x 3 mixed-design ANOVA with experimental group as a between-group variable and time of testing as a within-group variable was performed to analyze the scores on both science content tests. This was done independently for the knowledge, far generalization and abstract generalization items.

Additionally, a third set of analyses examined the accuracy of the assumptions underlying the experimental set up. It was assessed whether the misconception of the weight of an object being the most important factor for predicting sinking times held true in this sample. The prediction scores of both experimental conditions for the five experiments that violated the misconception that heavier objects always sink faster were compared by a one-way ANOVA with the 25 other experiments where weight was a legitimate explanation for differences in sinking times. The working of the prompts was also verified. Although supporting evidence evaluation skills with prompting can best be seen as a pooled intervention, it was looked into the working of an exemplary prompt to assess whether small changes in behavior took place. Before observing the third and sixteenth experiment, participants were prompted to repeatedly watch experiments because observing is not always that easy and repeating experiments is what scientists would do. It was examined whether a change in repetition behavior could be detected after following this prompt.

Results

The first set of analyses examined, based on logged user input and participants' notes, the mastery of evidence evaluation skills. Table 4 shows the number of evidence evaluation skills performed (prediction, observation, remarks on flaws due to authenticity, interpretation and generalization) of the full support, observation support and no support condition and also off-task remarks that could not be categorized as evidence evaluation skills. In all conditions, most logged user input and notes were relevant to the task, yet off-task remarks were highest in the full support condition. Predictions and observations were prompted in both experimental conditions in a way that there was no way to get past it. This also demonstrated itself in the number of predictions and observations. In the no support condition, the number of predictions that were written down spontaneously was very low, whereas the number of observations that were noted down without prompting constitute, averaged per participant, about half of the experiments. Remarks on flaws due to authenticity were highest in the no support condition. Interpretations and generalizations were only prompted in the full support condition. But even here, only about half of the notes contained statements that could be categorized as interpretation or generalization. It is remarkable that in the other conditions very few interpretations and generalizations were made voluntarily.

Table 4

Mastery of evidence evaluation skills, specified by the number of evidence evaluation skills performed (vs. off-task remarks) and the quality of this performance, derived from logged user input and notes, averaged per participant by experimental condition.

	Full support (n=15)		Observation support (n=17)		No support (n=16)	
	Number per participant	Mean correctness score (%)	Number per participant	Mean correctness score (%)	Number per participant	Mean correctness score (%)
Predictions	30 ^a	65.78	30 ^a	75.69	1.44	86.96
Observations	30 ^a	83.78	30 ^a	84.71	15.31	89.38
Remarks on flaws due to authenticity	0.33	-	0.59	-	2.06	-
Interpretations	16.87 ^a	22.04 ^c	0.88	1.23 ^c	6.44	10.69 ^c
Generalizations	2.73 ^b	45 ^d	0.24	7.35 ^d	0.69	14.06 ^d
Off-task remarks	2.6	-	0.06	-	0.125	-

Note. The total number of experiments was 30.

^a Prompted every experiment (30 times in total).

^b Prompted every fifth experiment (5 times in total).

^c From 62 points possible (Up to 4 attributes were correct per interpretation).

^d From 4 points possible (There were 4 attributes).

The quality of evidence evaluation skill performance is also displayed in Table 4. The mean correctness score for predictions and observations in the full support and observation support condition could be statistically analyzed, because they were made compulsory through prompting. Results show that both experimental groups did not differ significantly in how many correct predictions (full support condition: $M = 19.73$, $SD = 5.12$; observation support condition: $M = 22.71$, $SD = 4.33$; $F(1, 30) = 3.17$, $p = .085$) and correct observations were made (full support condition: $M = 25.13$, $SD = 2.82$; observation support condition: $M = 25.41$, $SD = 3.76$; $F(1, 30) = .05$, $p = .816$). The predictions and observations written down by the no support condition were also high in terms of correctness. Learners in the full support condition succeeded in achieving about 25 percent of all interpretation points and almost 50 percent of the generalization points. The other conditions lagged far behind these scores, in particular the observation support condition.

The mean number of repeated experiments was 1.67 ($SD = 2.16$; twice: $M = 0.20$, $SD = .41$) in the full support condition, 0.71 in the observation support condition ($SD = .85$; twice: zero) and 14.38 ($SD = 5.57$; twice: $M = 8.00$, $SD = 6.76$) in the no support condition. A one-way ANOVA revealed that differences in how often participants repeated experiments were statistically significant, $F(2, 45) = 77.59$, $p = .000$ (twice: $F(2, 45) = 21.80$, $p = .000$). A Bonferroni post-hoc comparison showed that participants repeated significantly fewer experiments in the full support condition (95% CI[9.60, 15.81]) and in the observation support condition (95% CI[10.66, 16.68]) than in the no support condition. They also repeated significantly fewer experiments twice in the full support condition (95% CI[4.30, 11.29]) and observation support condition (95% CI[4.61, 11.39]) than in the no support condition. Learners in both experimental conditions were prompted to repeat experiments which were difficult to observe; learners in the no support condition however also repeated many experiments that were clearly to discern.

Log files revealed that, on average, participants in the full support condition needed 23.82 minutes to complete the inquiry task ($SD = 9.06$). The mean time in the observation support condition was 15.99 minutes ($SD = 3.68$) and in the no support condition 25.69 minutes ($SD = 9.17$). A one-way ANOVA showed that the differences in time were statistically significant, $F(2, 45) = 7.51$, $p = .002$. A Bonferroni post-hoc comparison indicated that participants needed significantly less time in the observation support condition (95% CI[14.09, 17.89]) than in the full support condition (95% CI[18.79, 28.84]), or the no support condition (95% CI[20.81, 30.58]). Differences between the full support and the no support condition were not statistically significant at the .05 level.

The second set of analyses assessed whether a gain in science content was achieved by supporting the process of evidence evaluation. A mixed-design ANOVA was performed to analyze how participants' science content evolved from pre- to posttest in each experimental condition. Mean scores and standard deviations are displayed in Table 5. The overall score on the knowledge items indicated that participants had mediocre prior knowledge of sinking objects: the entire sample mean

was 10.67 point out of the 21.25 points possible. The mixed-design ANOVA produced no significant within-subject effect of time, $F(1, 90) = .08, p = .780$, no between-subject effect of experimental condition, $F(2, 90) = 1.52, p = .225$, and a non-significant time \times condition interaction, $F(2, 90) = .64, p = .528$. Comparable results were found for the generalization items. Neither a main effect of time, $F(1, 90) = 1.15, p = .287$, nor of condition, $F(2, 90) = .63, p = .534$, nor a significant time \times condition interaction $F(2, 90) = .33, p = .723$, could be detected for the scores on these items. Likewise, the scores on abstract generalization items showed no effect of time, $F(1, 90) = 1.31, p = .256$, condition, $F(2, 90) = .33, p = .720$, and no significant interaction, $F(2, 90) = .51, p = .601$. Together these findings indicate that participants in all three conditions had comparable pre-test and posttest scores, and no significant gains in science content.

Table 5

Mean pretest and posttest scores on the science content test by knowledge, far generalization and abstract generalization items and by experimental condition.

		Knowledge items			Far generalization items			Abstract generalization items		
		Full support (n=15)	Observation support (n=17)	No support (n=16)	Full support (n=15)	Observation support (n=17)	No support (n=16)	Full support (n=15)	Observation support (n=17)	No support (n=16)
Pretest	<i>M</i>	9.70	10.49	11.53	1.10	0.97	1.09	2.13	1.91	2.34
	<i>SD</i>	1.58	2.84	2.58	0.39	0.37	0.42	1.59	1.27	1.36
Posttest	<i>M</i>	10.29	11.09	10.80	1.20	1.12	1.09	2.07	1.74	1.63
	<i>SD</i>	2.83	2.63	3.37	0.32	0.42	0.33	1.39	1.35	1.27

Note. The maximum attainable score on the knowledge items was 21.25 points, on the far generalization items 1.5 points and on the abstract generalization items 7 points.

The third set of analyses examined the accuracy of the assumptions underlying the experimental set up. An analysis of the predictions made in both experimental conditions confirmed that the misconception about weight of an object being the most important factor for predicting sinking times does exist. Of the five experiments that violated this misconception, participants from the two experimental conditions had an average of 0.44 ($SD = 0.13$) of the predictions correct, whereas for the remaining 25 experiments they had an average of 0.77 ($SD = 0.11$) of the predictions correct. Performance on the critical experiments thus significantly differed, $F(1, 28) = 34.99, p < .001$, from performance on the other experiments. Figure 2 graphically displays the cumulated predictions of both experimental conditions.

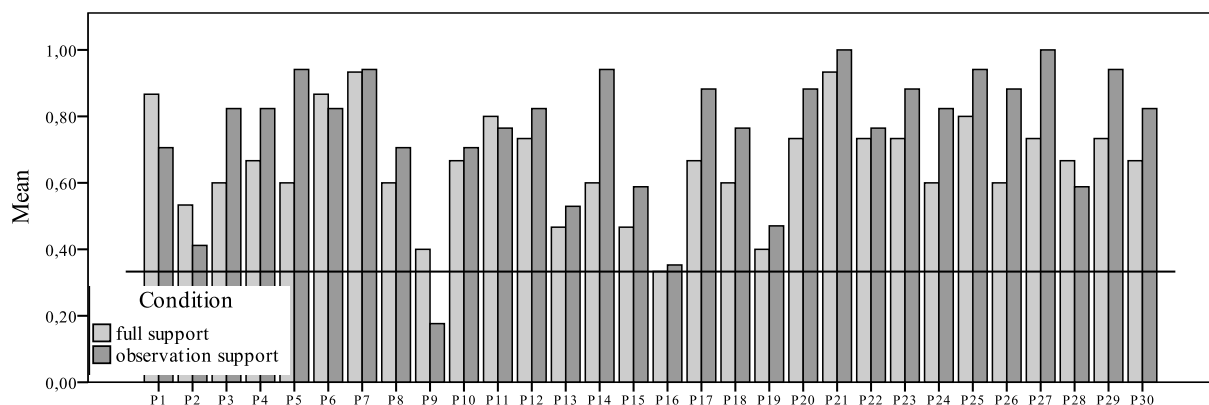


Figure 2. Cumulated predictions of the full support and observations support condition. At predictions P9, P13, P16, P19 and P28 the misconception that the weight of an object is the most important factor for predicting sinking times does not hold true. The black line represents the one-third chance of guessing.

As an exemplary demonstration of the effectiveness of the prompts, it was examined whether prompts could change repetition behavior. Directly following this prompt at the third and sixteenth experiment, 10 experiments were repeated, which was one-fourth of the repetitions participants in the experimental conditions made in total. Before the first of these prompts, just one repetition was made, which demonstrates the impact of this prompt.

Discussion

This study investigated whether process support can promote evidence evaluation skills and by that the acquisition of science content during inquiry learning. Overall results were disappointing. The mastery of evidence evaluation skills could only be promoted to some extent by the support, whereas no gain in science content could be detected.

Looking at evidence evaluation skills that were prompted, it becomes distinct that skill performance could only be affected to some extent. The correctness of observations for participants of both experimental conditions left room for improvement. The same is true for number and quality of interpretations and generalizations in the full support condition. Still, learners without full support were less successful at making interpretations and generalization. Although note taking was voluntary, from the notes participants of the observation support condition took, only very few could be categorized as interpretations and generalizations. In the no support condition more notes were taken but the number of them that could be categorized as interpretations and generalizations was by far not comparable to the full support condition. This is also reflected in the quality of interpretations and generalizations. Furthermore, it seems that learners who received prompts supporting the observation of evidence were more concentrated during the observations because they only repeated some experiments. Learners in the no support condition, by contrast, repeated an improper number of experiments - more than were reasonably necessary. Due to the effectiveness of the exemplary prompt, a basic impact of prompts can be evidenced.

Process support was also expected to enhance, as the result of a better performance of evidence evaluation skills, the acquisition of science content. This was what the bootstrapping idea predicted. The present study, however, could not find corroborating evidence. Seventh graders in both experimental conditions were equally unsuccessful in learning about sinking objects as pupils in the condition where no support was provided. No gain in science content could be attested for either condition on the knowledge items. The generalization of findings to other situations was equally unsuccessful in the experimental conditions as in the no support condition, as the results on the far and abstract generalization items show.

Based on Chinn and Malhotra's (2002a) conclusion that observation is the central skill to master for evidence evaluation to be effective and that too much support can have an reverse effect, it was expected that supporting observation skills only should lead to even better learning outcomes. However, scores on the science content test and prediction and observation scores in the observation support condition were comparable with those in the full support condition. The number and quality of interpretations and generalizations was even lower than in the no support condition. Making good observations might be the central activity of evidence evaluation, but it has not been confirmed that when the observation of evidence goes well also the rest does. Results showed that process support was less time consuming. Participants in the observation support condition needed less time than in the other two conditions. Compared to the full support condition, they made fewer off-task remarks, which seems to support the notion that too many prompts might overstrain learners and lead to a rejection of process support strategies. This was also confirmed by the fact that participants in the full support condition did only make meaningful interpretations and generalizations in half of the cases that they were prompted to do so.

Regarded as a whole, the results of this study do not confirm the expectations. However, two findings were quite noteworthy. It is remarkable how often learners in the no support condition repeated experiments. One possible explanation is that participants in the no support condition had great difficulties in observing evidence and that structure prompts succeed in scaffolding learners in the experimental conditions. Another interesting aspect is that pupils from the no support condition recorded more and better interpretations and generalization than pupils from the observation support condition. Perhaps supporting observation skills only has the undesired side effect of withdrawing the learners' attention from other evidence evaluation processes.

There are several possible explanations why process support did not enhance the acquisition of science content in this study. For one thing, the results give reason to question the quality of the science content test. Too many points could be scored by sticking to the misconception that heavier objects sink faster and there was too little opportunity to explain the influence of each attribute. The discriminating value of the science content test could be increased.

Process support prompts made learners engage in evidence evaluation skills, but, as far as this was measurable, could only enhance skill performance to some extent and did not lead to better learning of science content. This lets one assume that the engagement was only on a superficial level. Wichmann and Leutner (2009) made a good experience with extending support with regulative support that reinforces learners to pay attention to their own understanding by planning, monitoring, and evaluating their thought. Regulative support could supplement the basic approach of structuring, problematizing and exemplifying with prompts and hints for regulating learners' flow of thoughts.

The present findings once again confirm the robustness of misconceptions and children's reluctance to conceptual change. In this study, it was tried to support the process of observing, interpreting and generalizing from anomalous data to promote the restructuring of knowledge schemata. However, scores do not show any significant knowledge gains from pre- to posttest. Hence the prevailing misconception of the weight of an object being the most important factor for predicting sinking times did not change. As in Chinn and Brewer's (1998) and Lin's (2007) studies, learners may have rejected anomalous data and process support strategies could not prevent this.

The results of this research further exemplify what research in classroom settings can be like. Even with a controlled, well-designed experiment in which all possible precautions are taken, there are always some circumstances beyond the researcher's control. The motivation for participating in the inquiry activity was not optimal because of the lacking support from one teacher and the restraint of scheduling one experimental session in the last week of school before summer vacation. Of the 57 participants, 9 had to be removed from the sample. The high standard deviations in all measured variables further point at great differences among participants. This might in part be due to the fact that the study was conducted in a comprehensive school where the range from low to high achieving learners is naturally wide. However, high variations in scores can also result from differing motivations. It can anyhow be concluded that high standard deviations limit the conclusions that may be drawn about the intervention's effectiveness.

The results finally imply that the bootstrapping idea might not prove true. It is questionable whether a more sophisticated set of inquiry skills always promotes the generation of a deeper, more organized knowledge base. It cannot be decidedly determined whether the results of this research are due to the too low discriminating value of the science content test, the ineffectiveness of the intervention, or the fact that the assumption that promoting evidence evaluation skills leads to a better learning of science content does not hold true. In line with earlier attempts, see for example Lazonder and Kamp (2011), this study cannot decidedly show that an improvement in inquiry skills does lead to better content knowledge. However, that the bootstrapping idea holds the other way around (supporting the acquisition of domain knowledge leads to better skill performance) has been demonstrated in a number of studies. Lazonder et al. (2010), for example, found that learners that received content support before and/or during the inquiry activity conducted fewer exploratory experiments and instead generated more specific hypotheses.

Future research is needed to shed more light on these issues. When conducting similar research, one should make sure that the science content test has more discriminating power. The effectiveness of process prompts might be enhanced by complementing it with regulative support. Another recommendation concerns the topic of inquiry. When misconceptions exist, process support should focus more on promoting the acceptance of anomalous data in order to facilitate conceptual change. The study should also be replicated in another environment and with a larger, more homogeneous sample, to reduce within group variability. To clear away doubts about the bootstrapping idea, it should be tried to design a study that can better measure the intermediate step of the bootstrapping idea. Are participants actually better at performing evidence evaluation skills because of the support? It was unfortunately only to a very small extent possible to measure improvement with the present research design. The difficulties in measuring improvements in evidence evaluation skills were mostly linked to the choice of using prompts as a support intervention. Thus due to the circumstances that "asking questions might also scaffold the learning process and thus influence the exact processes one wants to study" (Wilhelm & Beishuizen, 2004, p. 251). Asking what participants have observed was an instance of a prompt in the experimental conditions. Also asking participants in the no support condition would however been the only way to get to know what they think and to compare skill performance.

Future research could be further expanded by adding a third measurement point after a couple of weeks to assess whether the restructuring of content schemata is permanent. If it is possible to let participants join in a series of supported inquiry activities, then it should also be evaluated at this point whether they retain skills. The goal is that learners cannot only perform a task beyond their current skills but also learn from this experience (Reiser, 2004). Such an assessment was beyond the purpose of the present study, but could be used as an ultimate test of an intervention's effectiveness to promote skill development and by that the learning of science knowledge.

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.
- Beishuizen, J., Wilhelm, P., & Schimmel, M. (2004). Computer-supported inquiry learning: effects of training and practice. *Computers & Education, 42*(4), 389-402.
- Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: a theoretical framework and implications for science instruction. *Review of Educational Research, 63*(1), 1-49.
- Chinn, C. A., & Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching, 35*(6), 623-654.
- Chinn, C. A., & Brewer, W. F. (2001). Models of data: a theory of how people evaluate data. *Cognition and Instruction, 19*(3), 323-393.
- Chinn, C. A., & Malhotra, B. A. (2002a). Children's responses to anomalous scientific data: how is conceptual change impeded? *Journal of Educational Psychology, 94*(2), 327-343.
- Chinn, C. A., & Malhotra, B. A. (2002b). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education, 86*, 175-218.
- Dewey, J. (1995). Science as subject-matter and as method. *Science & Education, 4*(4), 391-398. (Reprinted from *Science, 31*(787), pp. 121-127, 1910)
- Eberbach, C., & Crowley, K. (2009). From everyday to scientific observation: how children learn to observe the biologist's world. *Review of Educational Research, 79*(1), 39-68.
- Fund, Z. (2007). The effects of scaffolded computerized science problem-solving on achievement outcomes: a comparative study of support programs. *Journal of Computer Assisted Learning, 23*(5), 410-424.
- Kalyuga, S., Ayres, P., Chandler, P. & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*(1), 23-31.
- Kirschner, P., Sweller, J., & Clark, R. (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*(2), 75-86.
- Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Cambridge, MA: MIT Press.
- Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review, 96*(4), 674-89.

- 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.
- Lazonder, A. W. & Kamp, E. (2011). *Splitting up the inquiry question to promote children's scientific reasoning*. Manuscript submitted for publication, University of Twente, the Netherlands.
- Lazonder, A. W., Wilhelm, P., & Hagemans, M.G. (2008). The influence of domain knowledge on strategy use during simulation-based inquiry learning. *Learning and Instruction, 18* (6), 580-592.
- Lehrer, R., Schauble, L., & Lucas, D. (2008). Supporting development of the epistemology of inquiry. *Cognitive Development, 23*(4), 512-529.
- Lin, J. (2007). Responses to anomalous data obtained from repeatable experiments in the laboratory. *Journal of Research in Science Teaching, 44*(3), 506-528.
- Mulder, Y. G., Lazonder, A. W. & de Jong, T. (2010). Finding out how they find it out: An empirical analysis of inquiry learners' need for support. *International Journal of Science Education, 32*(15), 2033-2053.
- Penner, D. E., & Klahr, D. (1996). The interaction of domain-specific knowledge and domain-general discovery strategies: a study with sinking objects. *Child Development, 67*, 2709-2727.
- Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *Journal of the Learning Sciences, 13*(3), 273-304.
- Richtering, C. (2010). *Literature review*. Unpublished manuscript, University of Twente, Enschede, the Netherlands.
- Schauble, L. (1990). Belief revision in children: the role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology, 49*(1), 31-57.
- Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology, 32*(1), 102-119.
- Sodian, B. (1998). Wissenschaftliches Denken. In D.H. Rost, *Handwörterbuch der Pädagogischen Psychologie* (pp. 566-570.) Weinheim: Beltz.
- Vosniadou, S., & Brewer, W.F. (1987). Theories of knowledge restructuring in development. *Review of Educational Research, 57*, 51-67.
- Wichmann, A., & Leutner, D. (2009). Inquiry learning: Multilevel support with respect to inquiry, explanations and regulation during an inquiry cycle. *Zeitschrift für Pädagogisch Psychologie, 23*(2), 117-127.
- Wilhelm, P., & Beishuizen, J. (2004). Asking questions during self-directed inductive learning: effects on learning outcome and learning processes. *Interactive Learning Environments, 12*(3), 251-264.
- Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*(2), 172-223.

Appendix A

Example of an knowledge item from the science content test (pairwise comparison)



Both objects pictured below will be dropped simultaneously in cylinders filled with water.

Object	 Sphere small	 Cuboid small
Material	Glass	Stainless steel
Dimensions	Diameter $\phi=1$ cm	1,2 x 0,8 x 0,5cm
Volume	0,52cm ³	0,48cm ³
Weight	1,31g	3,84g

Which object will reach the bottom first?

Example of a far generalization item from the science content test

Both objects pictured below will be dropped simultaneously in cylinders filled with water.

Object	 Cylinder large	 Cylinder small
Material	Plastic (PET)	Plastic (PET)
Dimension	Radius $r=0,5$ cm; height $h=1$ cm	Radius $r=0,4$ cm; height $h=0,8$ cm
Volume	0,79cm ³	0,4cm ³
Weight	1,02g	0,52g

Which of these objects will reach the bottom first?

Example of a far generalization item from the science content test

Both objects pictured below will be dropped simultaneously in a swimming pool filled with water.

Which objects will reach the bottom first?

Why?

