Bachelor thesis Psychology
Faculty of Behavioural Sciences

UNIVERSITEIT TWENTE.

The relation between metacognitive skillfulness, intellectual ability, prior knowledge, causal mechanisms and inquiry learning performance in fourth- and sixth-graders

Name: Wiebke F. Evers

Student–ID: s0157104

First tutor: Dr. Pascal Wilhelm

Second tutor: Dr. Ard Lazonder

Date: October 26th, 2009

Abstract

Background. Inquiry learning offers learners the opportunity to construct knowledge on their own. By means of conducting simple experiments, the learner can test hypotheses about how certain variables influence each other. The obtained evidence then needs to be interpreted in a valid manner.

Aim. Different factors are assumed to influence the learning process. The first aim of this study was to investigate the effects of metacognitive skillfulness, intellectual ability, prior knowledge and the generation of causal mechanisms on inquiry learning performance. The second aim was to explore possible age-related differences in these effects.

Method. Fourth- and sixth graders were randomly assigned to two tasks in either the biology or the geography domain. Think-aloud protocols including responses to specific prompts were obtained to measure metacognitive skilfulness and the generation and use of causal mechanisms. Correlational analyses and MANOVA's were used to analyse the relationships between metacognitive skills, intellectual ability, prior knowledge, the generation of causal mechanisms and learning performance.

Results. The comparison between adults and children showed that scientific reasoning develops with age. The correlational analysis showed that the inquiry learning measures that are of importance in the learning process are consistent among children and adults. Causal mechanisms were significantly more generated by younger students. The generation of causal mechanisms showed to be negatively correlated with metacognitive skillfulness for the older students. A significant correlation was found between intellectual ability and metacognitive skillfulness. The effects of metacognitive skillfulness, prior knowledge and intellectual ability on learning performance depend on age.

Conclusion. Further research should consider the use of a regression analysis to obtain more information about the way metacognitive skills, intellectual ability, prior knowledge and the generation of causal mechanisms affect learning performance.

Table of Contents

1. Introduction

Over the past years, a vast amount of research has been conducted in the field of inquiry learning. Many of these studies aimed to find out more about the influence of metacognitive skillfulness, prior domain knowledge, and intellectual ability on learning performance in inquiry learning tasks (Veenman, Prins & Elshout, 2002). Metacognitive skillfulness and intellectual ability have been identified to be important predictors of learning performance (Van der Stel & Veenman, 2007; Veenman et al., 2002; White & Shimoda, 1999). The role of prior domain knowledge in inquiry learning however is still unclear. Both positive and negative effects on learning performance have been found.

Research has also shown that the generation of causal mechanisms influences the way the relationship between two variables is interpreted (Koslowski, 1996). The main objective of inquiry learning is to find out how two variables relate to each other. Nevertheless, the impact of the generation of causal mechanisms in an inquiry learning context has barely been investigated. Many questions about their role remain unanswered. Moreover, complex relationships seem to exist between the generation of causal mechanisms, metacognitive skillfulness, intellectual ability, and prior domain knowledge, influencing their individual effects on inquiry learning performance.

The aim of the present study is to contribute to the issue of how metacognitive skillfulness, intellectual ability, and prior knowledge are related to learning performance in inquiry learning tasks. Furthermore, the aspect of causal mechanisms will be investigated, specifically their relation to metacognitive skillfulness, intellectual ability, and prior knowledge as well as their effects on learning performance.

1.1 Inquiry learning

Inquiry learning is an educational approach that provides students with the opportunity to engage in an authentic learning experience in order to construct knowledge (Löhner, van Joolingen, Savelsbergh & van Hout-Wolters, 2005). Inquiry learning is an active and self-directed learning process: Learners conduct experiments on their own in order to find out how variables of a certain domain affect each other (Anderson, 2002). Over the past years, computer simulations made inquiry learning much more feasible (Löhner et al., 2005). They do not only offer interactivity and the possibility to manipulate objects (Rieber & Parmley, 1995) but also provide support to students lacking scientific reasoning skills. This aspect is

important because, although inquiry learning can be largely regarded as an effective way of learning, research has shown that not every student holds the cognitive skills that are needed for successful inquiry learning (Van Joolingen, de Jong, & Dimitracopoulou, 2007).

The processes that are evident in inquiry learning can be derived from the 'inquiry circle' of scientific research. The circle incorporates four transformative processes that enable students to generate new domain knowledge (e.g. De Jong & Van Joolingen, 1998; Njoo & De Jong, 1993; Saab, van Joolingen & van Hout-Wolters, 2005). The four processes are orientation, hypothesis generation, hypothesis testing (through experimentation), and drawing a conclusion. During the orientation process, the learner examines the variables presented in the task and gathers information about them. In the second phase, the learner comes up with a hypothesis on how the variables could relate to each other. Possible alternatives to the proposed hypothesis are formulated or ruled out. In the third process, the hypothesis is tested. A certain outcome may be predicted based on the hypothesis formed in the former phase. The last step of the inquiry circle is to draw a conclusion on the basis of the result of the experiment. The hypothesis is either confirmed or rejected by the experiment and the inquiry cycle starts over (Saab et al., 2005).

In order to be successful, the transformative processes need to be controlled by regulative processes such as planning, monitoring and evaluating (Van Joolingen, de Jong, Lazonder, Savelsbergh, & Manlove, 2005). Regulative processes are the major driving force behind the inquiry learning process (Njoo & De Jong, 1993). In the planning phase, the learner sets up a strategy of how to proceed in the inquiry learning process. The strategy should include goals and sub-goals to be reached by following a certain strategy. While conducting the experiments, the learner should monitor his progress in order to adjust his strategy if necessary. In the evaluation phase the learner can reflect on his learning progress and apply his findings to the problem he is presented with.

1.1 Metacognitive skillfulness

The term metacognition refers to the knowledge about one's own cognitive processes (metacognitive knowledge) as well as to the regulation of these processes (metacognitive skills) (Veenman, Prins, & Elshout, 2002). As metacognitive knowledge, which is the knowledge about the task characteristics and the applicable strategies, is covered by the term prior knowledge, the focus of this study has been laid on the latter, metacognitive skills.

There exist conflicting opinions on which skills are included in the concept of metacognitive skillfulness (Schneider, 2009). A vast amount of research on the role of metacognitive skillfulness has been conducted by Veenman et al. (2002; 2004; 2007). He identified four indicators for metacognitive skillfulness: orientation, systematic orderliness, evaluation and elaboration. The aspect of orientation, which means that the learner lays its focus on the relevant information of the task description, was not measured in the present study. Conducting the task systematically means that the learner is able to come up with a plan and sticks to it. The evaluation of its own progress is necessary for the detection of errors in reasoning the reasoning process. Elaboration means that the learners draws conclusions about the results and tries to find explanations for them. It could be noted that the term metacognitive skillfulness entails, with only small variations, the same cognitive processes as the term regulative processes, described above. During this study, the term metacognitive processes will be used.

Metacognitive skills play an important role in inquiry learning as they have a major influence on learning behavior (Van Joolingen, de Jong & Dimitracopoulou, 2007). Research has shown that it serves as a good predictor of learning performance (e.g. Van der Stel & Veenman, 2007; Veenman et al., 2002; White & Shimoda, 1999). High metacognitive skillfulness leads to good learning behavior, whereas learners with poor insight in their own learning process have shown poor learning performance (Veenman, Prins, & Elshout, 2002). Lack of metacognitive skills has been found to be quite common as an experiment conducted by Löhner et al. (2005) showed. They found out that, although students spend the majority of time on scientific reasoning activities, they often show deficiency in following a systematic approach.

1.2 Intellectual ability

Intelligence can be defined in many different ways. In the present study the rather pragmatic definition of Elshout (1983) is followed. He regards intelligence as a human cognitive toolbox enclosing basic cognitive operations. The magnitude and quality of the toolbox differs from person to person because of different biological influences such as genetic material and environmental factors such as school (Van der Stel & Veenman, 2007). Intellectual ability is seen as a major predictor of learning as it has shown to be positively correlated to the academic success of psychology students (e.g. Busato, Prins, Elshout & Hamaker, 2000). In a study by Veenman et al. (2002), intellectual ability has also been found

to be positively correlated to learning performance on an inquiry learning task. However, this correlation was only moderate. Therefore, it might be assumed that the influence of intellectual ability depends on the kind of learning that is studied.

There is some controversy on the relationship between intellectual ability and metacognitive skillfulness. Three different models exist that describe the relation between intellectual ability and metacognition (Veenman, Beishuizen, & Niewold, 1997). The first model is called the intelligence model and regards metacognition as a part of intellectual ability. As a result, metacognition is not able to predict learning performance on its own as it depends on intellectual ability. In the contrasting model, metacognition and intellectual ability are seen as two independent predictors of learning with no correlation existent between them. This model is supported by van der Stel and Veenman (2007). The third model is called the mixed model and combines both assumptions of the models described earlier. The model states that metacognition and intellectual ability combine to predict learning performance, but that each factor also contributes separately to this prediction (Veenman et al., 2004).

So, although there is some controversy about the way intellectual ability is related to metacognitive skillfulness, the effect of metacognitive skillfulness on learning performance in inquiry learning tasks is clearly a positive one.

1.3 Influence of prior knowledge

There are two types of prior knowledge that are assumed to have an influence on learning behavior (De Jong et al., 2005). The first one is prior domain knowledge, which enables the learner to generate relevant hypotheses (Van Lieburg, 2006) and the initial model the learner will have of the task as well as the design of the experiments the subject will conduct. The second type of knowledge that has an impact on learning performance is prior process knowledge (Van Joolingen, de Jong, & Dimitrakopoulou, 2007). Prior process knowledge is for example knowledge about inquiry learning skills which possession obviously has a positive influence on learning performance. As this knowledge is inherent to metacognitive skillfulness, the influence of prior process knowledge will not be discussed further.

The role that prior domain knowledge plays in inquiry learning is not clear yet. Prior knowledge can influence the learning performance positively as well as negatively. One the one hand, lack of sufficient prior knowledge is assumed to be an important reason for low inquiry learning performance (Lazonder, 2008). Possessing process knowledge on how to

perform inquiry learning processes will obviously influence the performance on the task positively (Van Joolingen, de Jong & Dimitrakopoulou, 2007). On the other hand, high levels of prior domain knowledge can limit the learners' view on the problem by leading them in a certain direction of experimentation. Moreover, the interpretation of the obtained evidence might be hindered as only evidence consistent with the learners' prior knowledge is taken into account (Kuhn, 2008).

Klahr and Dunbar (1988) developed the Scientific Discovery as Dual Search model (SDDS model) that offers an explanation for the role prior knowledge in inquiry learning. The model assumes that the strategy that is used by the learner depends on their prior knowledge. The inquiry process of a learner can be seen as the search through two different spaces: the hypothesis space and the experiment space. The student's knowledge about the relations between the variables in the domain as well as the knowledge that is accumulated during the experiment is present in their hypothesis space. The hypothesis space is searched by the learner for relevant hypotheses. The experiment space on the other hand, contains all experiments that can be conducted with the learning materials available. The hypotheses that have been generated earlier can give direction to the search through the experiment space and lead to valid experiments to test the hypotheses. The evidence that has been collected through the experiments then is evaluated in order to verify or refine the hypotheses.

Prior domain knowledge has a major influence on the search in the hypothesis space which determines the strategy the learner will follow. Research conducted by Klahr and Dunbar (1998) has shown that learners can be divided into two groups, theorists and experimenters, based on their strategy use. The theorists and the experimenters both start by searching the hypothesis space and then conduct experiments to test their hypotheses. If the hypothesis does not seem to fit the obtained evidence, the theorists revise the hypothesis by taking their prior domain knowledge into consideration. The experimenters on the other hand, only posses a limited amount of prior knowledge from which they can set up hypotheses. If they are not able to make new hypotheses on the basis of their prior knowledge or the outcomes of the already conducted experiments, they switch from searching the hypothesis space to searching the experiment space, where they carry out explorative experiments. The experiments are not based on explicit hypotheses but should enable the learner to obtain enough knowledge to switch back to the hypothesis space. Klahr and Dunbar (1998) found out that theorists seem to reach higher learning performance scores than experimenters. So, it seems that prior knowledge has a positive influence on the scientific reasoning process and therefore on learning performance.

Lazonder (2008) corroborated Klahr and Dunbar's model. He investigated how prior domain knowledge influences investigative strategies of learners. Participants with a high level of prior knowledge engaged in a theory-driven strategy, whereas participants with lower levels of prior knowledge employed a more data-driven mode of inquiry as they lack sufficient domain knowledge to set up relevant hypotheses. In the course of the experiment, they switched to the theory-driven strategy as more knowledge is obtained through the conducted experiments. Lazonder's findings support the SDDS model of Klahr and Dunbar as they confirmed the existence of two different approaches based on the level of prior knowledge of the learner.

In contrast to the results obtained by Lazonder, Hulshof (2001) found no significant difference in performance on an inquiry learning task between subjects with high and low levels of prior knowledge. Research by Kuhn (2008) showed that higher levels of prior knowledge could even weaken the subject's performance. As it was stated before, prior knowledge can influence the generation of an initial theory about the way the input and output variables relate to each other, which was assumed to benefit learning performance. However, research has shown that especially children often have trouble distinguishing between their generated theory and the obtained evidence with the result that evidence is not interpreted in a valid manner. Evidence that would falsify the child's theory is often rejected or ignored. This lack of valid coordination between theory and evidence offers an explanation for the qualitative difference between children and adults in scientific reasoning. In order to engage in successful inquiry learning, learners have to be able to disregard their prior knowledge and rely on the data present in the task (Kuhn, 1989).

To conclude, the influence of prior domain knowledge can be beneficial as well as hindering, dependent on how the prior knowledge is used by the learner.

1.4 Causal mechanisms

The term causal mechanism is defined as the additional information that is expressed by the learner in order to explain the influence of one variable on another. By additional information it is meant that the learner mentions information other than the information already given in the task. This could be for example a description of what intervenes or mediates between an input variable affecting an output variable. Causal mechanisms are often based on prior beliefs a person holds but are also generated or adapted spontaneously when a judgment about a relationship is required (Koslowski, 1986). Knowledge about

causal mechanisms can facilitate the identification of the best explanation for an assumed or observed relationship between two variables (Ahn & Kalish, 2000).

There are conflicting findings in research on the effect that causal mechanisms have on learning performance and how it is related to other evidence that is available to the learner about the relationship between the independent and dependent variables in the task. Koslowski (1996) stated that adults as well as three-year old children regard the presence of a causal mechanism to be more important than covariation when asked to make a statement about the relationship between two variables. Koslowski cites a study conducted by Kuhn, Amsel and O'Loughlin (1988) which gives support to the view that using causal mechanisms influences the evaluation of covariations significantly. Subjects were asked to investigate the influence of four variables (size, color, texture, and the presence and absence of ridges) on the bounceability of sports balls. Almost all subjects regarded the aspect of color as having no influence on the bounceability of the balls. When presented with evidence that color and texture covaried with bounceability, the covariational evidence for color was ignored. This finding showed that the causal mechanism stating that color has no effect is regarded as more important as the observed covariation between color and bounceability. To put it differently, covariation between variables is only regarded as evidence for causality if a causal mechanism exists that gives a logic explanation for the relation between the variables (Kuhn et al, 1988; Kuhn, 2004). Moreover, it has been found that if covariation turns up and the subject does not have any knowledge about a causal mechanism that could explain the covariation, the subject assumes that a causal mechanism exists that has just not yet been discovered. Hence, the pure existence of covariation does not seem to be enough for a subject to believe in a relationship between two variables. The presence of causal mechanisms that offers a reason to believe that the observed covariation is not just an arbitrary or coincidental one is much more important than the presence of covariation alone (Koslowski, 1996).

As said, causal mechanisms are based on the prior knowledge a person holds (Koslowski, 1996). Research has shown that prior knowledge can be beneficial as well as disadvantageous to the subject's learning performance. Consequently, the same controversy exists on the role of causal mechanisms. On the one hand, causal mechanisms are regarded as useful for evaluating correlations that occur in the world by using the knowledge we already have. Causal mechanisms can provide evidence when judging the relationship between two variables and foster the development of relevant hypotheses about it. On the other hand, this alleged evidence can also limit the view of a person with the result that facts that would

disconfirm their theory about how the two variables are related are ignored (Evans, Barston & Pollard, 1983).

The previous paragraphs might give the impression that causal mechanisms are generated frequently and are important for judging relations between events in the real world. However, research has shown that there are significant differences between the use of causal mechanisms in the real world and in experimental settings. A study conducted by Löhner et al. (2005) in which college freshmen were allocated to three different inquiry learning tasks showed that the use of causal mechanisms in experimental settings where inquiry learning tasks are used is uncommon. Causal mechanisms were used only by a small number of participants as a support of a hypothesis and that for the most formulated hypotheses no further explanations were given. A possible explanation for this low usage of causal mechanisms is age. It has been found that older subjects have shown to be more test-wise than younger subjects, as they realize that the inquiry learning setting demands them to ignore their knowledge of causal mechanisms. Koslowski (1996) argued that in a less artificial learning environment the older students would also rely on their knowledge of mechanisms. In other words, depending on the artificial judging situation learners are put in, their normal judgment behavior might be affected. The more artificial the experiment is, the less causal mechanisms will be generated.

Another explanation for the low usage of causal mechanisms in inquiry learning is tasks, is offered by Koslowski (1996). She found out that the generation of causal mechanisms in inquiry learning tasks decreased with age. Younger children have been found to use causal mechanisms more often to support their theories than deriving information from covariation evidence. Children state causal mechanisms to explain their decision to take some covariation evidence into account and others not. Letting go of a causal belief has shown to be more difficult for children that the relinquishment of a non-causal belief.

The use of causal mechanisms in inquiry learning tasks is often regarded as a flaw in scientific reasoning. In an inquiry learning task, the only source of evidence the subject should rely on is covariation (Koslowski, 1996). Deriving evidence from causal mechanisms or theories is often treated as being inferior as it shows poor scientific reasoning and low levels of metacognitive skillfulness. Disregarding causal mechanisms is seen as good scientific reasoning (Kuhn et al, 1988).

Koslowski (1996) however, argues that the reliance on causal mechanisms must be regarded as legitimate as it is a valid and useful aid for judgments in the real world. Causal mechanisms can display the mediator between two events and its presence can function as

evidence that an observed correlation is causal. By simply reducing causation to covariation, a distorted picture of the reasoning process of the subjects will be obtained.

Koslowski also mentions some shortcoming in the way that evidence based on a causal mechanism is utilized. She criticizes that subjects are not sufficiently critical about causal mechanisms they hold. The flaw in scientific reasoning is not the usage of causal mechanisms but the fault to ignore covariations that are implausible when evaluated on the basis of the causal mechanisms. These finding might also explain the trouble younger children face coordinating theory and evidence.

So far, the generation of causal mechanisms and the effect of its presence on learning performance in inquiry task have barely been investigated. In addition, little is known about how the use of causal mechanisms is related to metacognitive skillfulness, prior domain knowledge, and intellectual ability.

1.5 Research questions

The first objective of this study is to replicate the findings presented in Wilhelm (2001) with two younger age groups. Using a scoring protocol, Wilhelm examined the learning behavior of adults in two inquiry learning tasks. A correlational analysis between the obtained inquiry learning measures and comprehension score displayed several significant relations. By comparing the results obtained for the two age groups involved in the present study with each other as well as with the results found by Wilhelm (2001) for adults, age differences in inquiry learning can be studied. It is aimed to find out which inquiry learning measures show a strong correlation with learning performance. Although this research objective is to a great extent exploratory in nature, it is hypothesized that significant differences between the three age groups on the inquiry learning measures will be found.

In addition to the replication of findings on inquiry learning behavior with a younger age group, the second objective of the study is to investigate the effects of causal mechanisms, metacognitive skillfulness, intellectual ability, and prior knowledge on learning performance. Moreover, it is aimed to investigate how the generation of causal mechanisms, metacognitive skillfulness, intellectual ability, and prior knowledge affect each other. By comparing the results of the two age groups, possible age dependent effects can be investigated.

There has been evidence that causal mechanisms play an important role in a subject's judging process. In the real world, the presence of a causal mechanism that gives an explanation for the relationship between two variables was found to be more important than

an observed covariation. With the present study, we set out to investigate how the generation of causal mechanisms contributes to inquiry learning. Based on the findings of earlier conducted studies (Kuhn, 2004; 2008), it is hypothesized that causal mechanisms will affect learning performance on inquiry task negatively. The generation and the application of causal mechanisms in inquiry learning task are often regarded as weak scientific reasoning and low levels of metacognitive skillfulness. Consequently, it is hypothesized that causal mechanisms will show a negative correlation with metacognitive skillfulness. As little is known about the relations between the generation of causal mechanisms, prior knowledge and intellectual ability, this part will be exploratory. However, research found evidence for age related differences in the use of causal mechanisms. The generation of causal mechanisms in inquiry learning task was found to decrease with age. Therefore, it is hypothesized that the fourth-graders will generate significantly more causal mechanisms than the sixth-graders.

Earlier studies on inquiry learning indicated that metacognitive skillfulness is positively related to learning performance. Higher levels of intellectual ability were also found to benefit the student and result in higher scores on learning performance. Therefore, it is hypothesized that higher levels of intellectual ability would be positively correlated to learning performance. Hence, both metacognitive skillfulness and intellectual ability have been found to be good predictors of learning behaviour. Moreover, high levels of intellectual ability have been shown to be related to high levels of metacognitive skillfulness. Therefore, it is hypothesized that a positive relationship between the two concepts will be found. Little is known about the relation of metacognitive skillfulness and intellectual ability with prior knowledge. This part will consequently be exploratory.

As it has been said before, prior knowledge can affect learning performance in a positive as well as in a negative way. It was found that, with adults, prior knowledge had a positive effect on inquiry learning performance as it enabled the learner to generate relevant hypotheses. In the present experiment, tasks from two different domains (biology and geography) are used. The two domains are both covered by the school curriculum, so every student who participated in the study had some expertise in it. It is assumed that a more realistic task would allow the learners to make use of causal mechanisms that are based on their prior knowledge and therefore gain more confidence in their ideas and opinions. On the other hand however, the experimental setting of the study is obviously an artificial one. Therefore, the students may not apply their prior knowledge to the problem presented in the task as they know that the prior beliefs they hold might not fit the underlying model of the

task. As very little research has been conducted in this area so far, the statement of hypotheses remains difficult, so this objective is more exploratory.

2. Method

The data that is used in the present study has been obtained by Wilhelm (2001) for his dissertation. A part of the data was reported in Veenman et al. (2004). However, the thinking-aloud protocols of the two age groups included in this study have not been analyzed before. Moreover, the aspect of the use of causal mechanisms was added. To this end, the original protocol was extended. The method and procedure is only explained shortly as it can be read in greater detail in Wilhelm (2001) or Veenman et al. (2004).

2.1 Participants

Sixty-three students from an elementary school in the urban area of Amsterdam participated in the study. Due to tape-recording failures, ten of the 63 thinking-aloud protocols could not be analyzed. Accordingly, the sample size of the experiment was 53 students. One group consisted of 28 fourth-graders (mean age: 9.5 years, SD= 0.36) with 15 being female and 13 male. The second group comprised 25 sixth-graders (mean age: 11.62 years, SD= 0.47) with 12 of them being female and 13 being male.

The parents of the children involved in the study were informed in advance about the purpose of the experiment and had to give permission. As a reward for their participation in the study the students received candy bars.

2.2 Materials

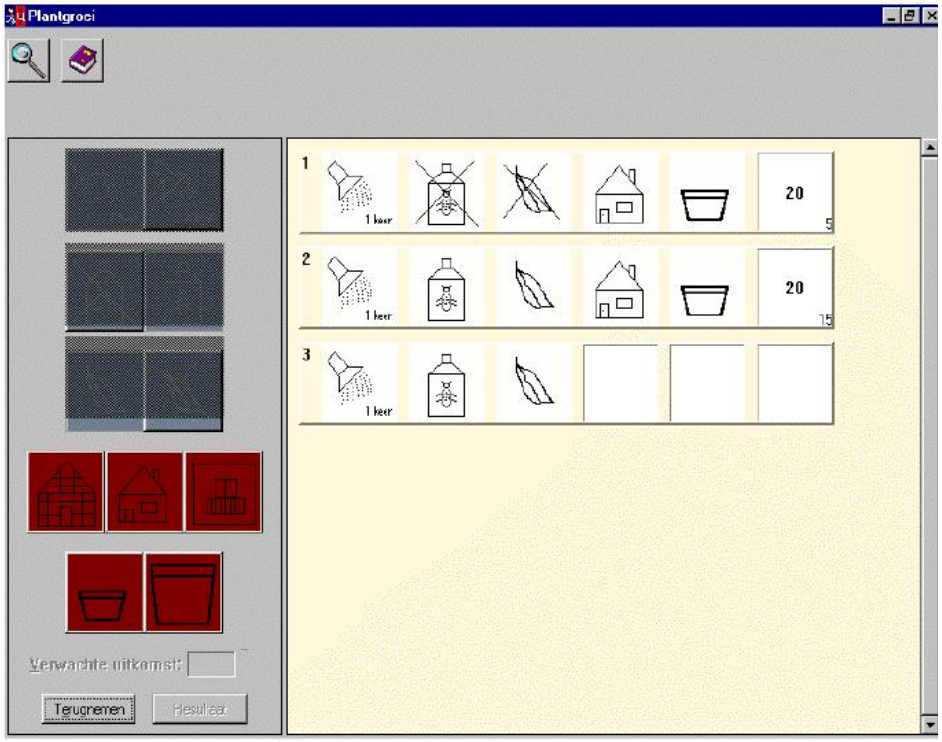
2.2.1 Inquiry learning task

In the experiment, four different computerized inquiry learning tasks, two from the domain of biology and two from the domain of geography, were used. Pilot testing indicated that children produced considerately more causal mechanisms in the first two tasks they conducted, so only the first two tasks the children did were used for analysis. Consequently, 14 fourth-graders and 12 sixth-graders conducted the two tasks of the biology domain; 13 fourth-graders and 12 sixth-graders did the two tasks of the geography domain. In the

biology domain the students did the plant task first and then moved on with the food task; in the geography domain they conducted the otter task first and then the ageing task.

The relationships between the variables in the tasks are in agreement with real-life phenomena. The underlying model describing the relationships between the independent and dependent variables is the same in each of the four tasks. Two of the five variables interact, one has a curvilinear effect (i.e. one of the levels of the variable had an effect, the others did not) and two variables are irrelevant to the outcome. An example of how the tasks were presented on the computer screen to the learners is displayed below.

Figure 1: Interface of the Plant task



With the help of a scorings protocol and the assessment of comprehension scores the learners' performance and behavior on the tasks were evaluated. The original scoring protocol developed by Wilhelm (2001) was designed to measure the learners' planning behavior, the generation of hypotheses, the testing of hypotheses and their drawing of conclusions of the experiments they conducted. For this study, the scoring protocol was extended to assess the frequency of causal mechanisms generated and the extent to which their presence affected experimental behavior. For the scoring protocol (in Dutch) see Appendix A.

At the end of each task, the learners have been asked to state his or her conclusions about the effect each variable had on the output variable. These statements were then compared to the model underlying the task and produced the comprehension score. The maximum comprehension score for each task was 18 points. The comprehension scores have been assessed for the original study and will be reused here. For more details on the scoring method see Wilhelm (2001).

2.2.2 Metacognitive skillfulness

Metacognitive skills are highly interdependent. It is assumed that already a small set of measurements are able to indicate the level of metacognitive skillfulness a subject holds (Veenman et al., 2002). In their study, Veenman, Wilhelm and Beishuizen (2004) used as indicators of metacognitive skillfulness log-file measures displaying the number of variables changed per experiment and the number of scrolling actions back to earlier conducted experiments. Changing more than one variable from one experiment to the next was assumed to represent poor systematic behavior. Following the Control-of-Variables Strategy (CVS, Chen & Klahr, 1999), which means the manipulation of only one variable from one experiment to the next, was assumed to indicate good systematic behavior as it enables the learners to draw valid conclusions from their experiments. High number of scrolling actions was assumed to indicate high levels of metacognitive skills. In order to verify the suitability of the two indicators obtained through the log-files, a protocol analysis of 15% of the protocols was conducted. The protocols were judged by two raters on the following four criteria: the quality of (1) orientation (elaborateness of hypotheses generated before each experiment); (2) systematical behavior (planning a sequence of experiments, and avoiding unsystematic behavior such as varying two independent variables between subsequent experiments); (3) evaluation (detection and correction of mistakes); and (4) elaboration (drawing conclusions, relating outcomes of experiments, generating explanations, and recapitulating). It was found that the measures were able to indicate metacognitive skillfulness (Veenman, Wilhelm & Beishuizen, 2004).

In the present study, the indicators of metacognitive skillfulness were chosen differently. Through the conduction of a protocol analysis of all students' think-aloud protocols, inquiry learning behavior was measured as in Wilhelm (2001). Those inquiry learning measures (indicative of metacognitive skillfulness) which showed a significant correlation with comprehension scores were selected. To obtain evidence whether these inquiry learning

measures were suitable to serve as an indicator for metacognitive skillfulness, a correlational analysis among them was conducted. Those measures showing the highest correlations with other indicators were picked and combined to one variable, indicating metacognitive skillfulness. This compound measure was used for the correlational analysis on metacognitive skillfulness and intellectual ability, prior domain knowledge, and causal mechanisms.

2.2.3 Intellectual ability

The tests that were used to assess the intellectual abilities of the learners were chosen on the basis of findings of Elshout (1976). He correlated scores on various (sub) tests for intellectual ability with their learning performance on an inquiry task. Six tests that were highly related to inquiry performance were selected: Number Series (Elshout, 1976), Concrete Syllogisms, Abstract Syllogisms (Conclusions, Elshout, 1976), Hidden Figures (Flanagan, 1951), Spatial Insight (DAT, Evers & Lucassen, 1983), and a test with mathematical word problems which included items from both Elshout (1976) and Klavier, Mommers, de Visser and Warners (1977). For more detailed information on the tests see Wilhelm (2001).

2.2.4 Prior knowledge

To test prior domain knowledge the learners held, two multiple choice tests were administered prior to the actual inquiry task. Each of the two tests contained three different sets of questions. The first ten questions of each test concerned the domain that was covered in the tasks. The learners were asked to state their knowledge about the five independent variables coming up in the task. The second part of each test consists of ten questions from the CITO test (CITO, 1998). The CITO test is usually taken by students in the last grade of elementary school to measure their scholastic aptitude and skills as well as their social and emotional development. The last set of questions consists of 20 questions that had been taken of the secondary school exam (VBO-MAVO; CITO, 1996-1997). By using questions regarding the biology and geography domain that are asked in the CITO and secondary

school exam, it was made sure that the questions were suitable for the learners across all wide age-range. The maximum score on each test was 40 points.

2.2.5 Causal mechanisms

The original scoring protocol was expanded to measure the generation of causal mechanisms. Four measures have been obtained with regards to causal mechanisms: the presence of causal mechanisms, the number of causal mechanisms made per experiment, the number of unique causal mechanisms, and the application of the causal mechanisms made. By application of the causal mechanisms it is meant, that the mechanisms that are stated have an influence on the hypothesis generation, experimentation or inferences that are made in the task. This comprehension would give evidence that the generated causal mechanisms affect the reasoning process and serve as a source of evidence for the learner.

2.3 Procedure

After the assessment of prior knowledge and intellectual ability of the learners, the individual testing sessions began. Each learner did both the biology and the geography tasks, although only the think-aloud protocols of the first two tasks they conducted were analyzed for this study. The learners were randomly assigned to two conditions which differed in whether they first performed tasks in the geography or the biology domain. The order in which the tasks were presented to the learner was always the same within each domain. Learners assigned to the biology condition did the plant growing task first followed by the food task. In the geography condition the learners started with the otter task and then moved on to the ageing task.

The learners executed the tasks in a separate room. They were placed in front of a computer with the experimenter sitting next to them. First, they were told that the purpose of the experiment was to find out more about the way people learn independently when no explicit instructions are given to them. Then the experimenter showed the learners an example of a task similar to the one they had to execute. They received a five-minute introduction into the program, learning how to change the levels of the independent variables, enter a prediction and run an experiment.

To keep track of the learners' reasoning, they were asked to think aloud while executing the experiment. A two-minute audio recording was played to the learners in order to demonstrate how to think aloud. A sensitive microphone attached to the computer recorded the verbal utterances made during the task.

The learners began the task with reading a story in which the problem was described (for an example see Appendix B). The icons representing the different levels of the independent variables were explained to them and the purpose of the task was repeated. Then, the experimenter tapped the learners' expectations about how each variable would affect the outcome. For each variable, the experimenter asked: "In what way, do you think, will (independent variable) make a difference for the (dependent variable)?" The data obtained through the questions of this pre-interview has not been evaluated in the present study. At the start of each experiment, the experimenter asked: "What are you going to find out?" and after the learners had entered a prediction and ran the experiment, they were asked: "What have you found out?" If the learners made a statement in relation to the effect of a specific variable they were asked by the experimenter: "Which one of the experiments that had been executed so far supports this conclusion?" The questions were asked in order to obtain specific information in addition to the think-aloud protocols. A study conducted by Wilhelm and Beishuizen (2004) has shown that these questions have no effect on learning outcome. The learners were also allowed to use a sheet of paper as a memory aid on which they could write down notes about their findings and conclusions.

When the learners expressed that they completed the task, they were asked to state precisely the effect that each of the independent variables had on the outcome. Using the scoring protocol, a comprehension score for each task was calculated. For more details on the procedure, see Veenman, Wilhelm & Beishuizen (2004).

2.4 Data-Analysis

Correlation analyses were used to study the relations between prior knowledge, intellectual ability, metacognitive skillfulness, causal mechanisms and comprehension score, indicating learning performance. To evaluate differences between the two age groups, multivariate analysis of variance was conducted.

3. Results

First of all, the percentages of agreement between the two raters for the different categories of the extended scoring protocol were assessed. One task was scored by the two raters together to calibrate the protocol and eight tasks conducted by four learners (three fourth-graders, one sixth-grader) were evaluated by both raters separately. Two of the learners were assigned to the biology and two to the geography condition. The mean percentage of agreement for the different categories was 82%. The rest of the protocols were evaluated by one rater only. The mean percentage of agreement on the comprehension score was 81% and was assessed earlier by Wilhelm (2001). It has not been assessed again for this study; the original data were used.

3.1 Comprehension Scores

The mean comprehension scores per age group are depicted in Table 1. Comparing the scores the learners obtained on both task, sixth-graders reached significantly higher comprehension scores than fourth-graders ($F(1, 52) = 18.30$, $p < .01$). Comparing the comprehension scores within each age group, a significant difference is found for the fourth-graders between the scores on the two domains. The fourth-graders assigned to the geography domain scored significantly higher than the fourth-graders assigned to the biology domain ($F(1, 54) = 5.43$, $p < .05$). No significant differences in comprehension score on separate tasks were found, indicating the absence of task-dependent effects.

***Table 1**: Means and Standard Deviations of Comprehension Scores in both domains and age groups*

	Grade 4: C-Score			**Grade 6: C-Score**		
	M	**SD**	**n**	**M**	**SD**	**n**
Plant growing task	6.27	2.81	14	7.08	3.34	12
Food task	7.73	3.20	14	9.42	4.27	12
Total Biology	6.71	2.94	14	8.25	3.94	12
Otter task	7.36	2.68	14	8.00	3.46	13
Ageing task	8.93	2.53	14	9.69	2.98	13
Total Geography	8.14	2.68	14	8.85	3.28	13
TOTAL	7.43	2.38	28	**9.30****	2.08	25

Note. C-Score: Comprehension score. Maximum score: 18 points.

3.2 Inquiry learning measures and learning performance

The inquiry learning measures obtained through the analysis of the scoring protocols were correlated to the comprehension scores. All measures were weighted by the number of experiments each learner conducted. For a definition of the measures, see Wilhelm (2001).

The analysis of the correlation between the overall comprehension score of both grades on the two domains and the inquiry learning measures leads to several significant results (see Table 2). Taking both groups together, significant correlations between overall comprehension score and the following measures were found: formulating hypotheses ($r = .42$, $p < .01$), formulating simple hypotheses ($r = .48$, $p < .01$), usage of CVS ($r = -.44$, $p < .01$), making valid inferences ($r = .40$, $p < .01$), the percentage of valid inferences ($r = .46$, $p < .01$), formulating plans ($r = .46$, $p < .01$), checking one variable ($r = .49$, $p < .01$), checking the effect of a variable in a different experiment ($r = .36$, $p < .01$), stating inductive plans ($r = .45$, $p < .01$), and following an undirected plan ($r = -.46$, $p < .01$).

For the fourth-graders, the following correlations were significant only: making valid inferences ($r = .38$, $p < .05$), the percentage of valid inferences ($r = .38$, $p < .05$), checking one variable ($r = .47$, $p < .05$), and stating inductive plans ($r = .38$, $p < .01$).

Evaluating the scores of the sixth-graders, the correlations between comprehension score and four inquiry learning measures turned out to be significant. These measures were formulating hypotheses ($r = .50$, $p < .05$), stating complex hypotheses ($r = .43$, $p < .05$), formulating plans ($r = .44$, $p < .05$), and following undirected plans ($r = -.44$, $p < .05$).

Table 2: *Correlations between Inquiry Learning Measures and Comprehension Scores across Domains: fourth- and sixth-graders*

	Total (n=53)	**Grade 4 (n=28)**	**Grade 6 (n=25)**
Hypothesis generation			
Formulate hypothesis	**.42****	.15	**.50***
Simple hypothesis	**.48****	.36	.39
Complex hypothesis	.02	-.30	**.43***
Design of experiments			
Coverage experiment space	.14	-.04	.07
CVS	**-.44****	-.37	-.34

*p < .05, **p < .01

Table 2 (continued)

	Total (n=53)	Grade 4 (n=28)	Grade 6 (n=25)
Interpretation of data			
Formulate inference	.10	.04	.15
Valid inference	**.40****	**.38***	.12
Percentage valid inferences	**.46****	**.38***	.21
Prediction error	.04	.09	.04
Regulation of Learning			
Formulate plan	**.46****	.36	**.44***
Checking one variable	**.49****	**.47***	.32
Checking interaction	-.01	-.15	.23
Generate specific outcome	-.06	-.07	.07
Check effect of variable in different experiment	**.36****	.29	.38
Predict outcome	-.16	-.25	.11
Inductive plans	**.45****	**.38***	.37
Deductive plans	.11	-.04	.31
Undirected plans	**-.46****	-.36	**-.44***
Causal mechanism			
Generation of causal mechanisms	-.19	-.16	-.04
Application of causal mechanisms	-.02	.02	-.03

*p < .05, **p < .01

A comparison of the inquiry learning measures across domains displayed several differences between the two age groups as shown in Table 3. The sixth-graders reached significantly higher scores on formulating hypotheses ($F (1, 51) = 10.96$ $p < .01$), formulating simple hypotheses ($F (1, 51) = 4.34$, $p < .05$), coverage of the experiment space ($F (1, 51) = 5.47$, $p < .05$), the generation of valid inferences ($F (1, 51) = 50.97$, $p < .01$) as well as the percentage of valid inferences made ($F (1, 51) = 126.03$, $p < .01$). They also scored higher on the formulation of plans ($F (1, 51) = 4.44$, $p < .05$), checking one variable ($F (1, 51) = 5.44$, $p< .05$), and stating inductive plans ($F (1, 51) = 4.99$, $p < .05$). The fourth-graders scored significantly higher than the sixth-graders only on the use of undirected plans ($F (1, 51) = 4.44$, $p < .05$).

Table 3: *Means and Standard Deviations of Inquiry Learning Measures across domains: Grade 4 and 6*

	Total (n=53) **M (SD)**	**Grade 4 (n=28)** **M (SD)**	**Grade 6 (n=25)** **M (SD)**
Hypothesis generation			
Formulate hypothesis	.25 (.16)	.19 (.15)	**.33 (.16)****
Simple hypothesis	.21 (.14)	.15 (.12)	**.28 (.14)***
Complex hypothesis	.04 (.06)	.04 (.07)	.05 (.05)
Design of experiments			
Coverage experiment space	23%	22%	**25%***
CVS	2.19 (.47)	2.25 (.48)	2.01 (.46)
Interpretation of data			
Formulate inference	1.02 (.39)	.93 (.39)	1.11 (.37)
Valid inference	.63 (.29)	.30 (.23)	**.97 (.35)***
Percentage valid inferences	58%	31%	**88%***
Prediction error	1.02 (.18)	1.02 (.21)	1.01 (.17)
Regulation of Learning			
Formulate plan	.45 (.25)	.38 (.24)	**.52 (.25)***
Checking one variable	.28 (.20)	.23 (.22)	**.35 (.16)***
Checking interaction	.04 (.08)	.04 (.10)	.04 (.07)
Generate specific outcome	.06 (.11)	.07 (.13)	.04 (.09)
Check effect of variable in different experiment	.05 (.07)	.04 (.06)	.06 (.08)
Predict outcome	.03 (.06)	.03 (.08)	.03 (.04)
Inductive plans	.32 (.23)	.26 (.24)	**.39 (.18)***
Deductive plans	.12 (.13)	.12 (.14)	.13 (.13)
Undirected plans	.55 (.25)	**.62 (.24)***	.48 (.25)
Causal mechanism			
Generation of causal mechanisms	.32 (.27)	.39 (.27)	.25 (.26)
Application of causal mechanisms	.10 (.10)	.11 (.09)	.10 (.10)

Note. Measures are weighted by number of experiments conducted. *p < .05, **p < .01

It was aimed to combine the inquiry learning measures into one variable capable of representing metacognitive skillfulness. As it is displayed in Table 2, ten measures showed a significant correlation with comprehension score. However, the generation of valid inferences, as opposed to the mere presence of an inference, not only indicates evaluative behavior as such but also assesses the quality of the evaluation that has been made. Therefore, the two measures valid inferences and percentage of valid inferences, which also displayed significant correlations with comprehension score, were not involved.

Among the eight other factors, a correlational analysis has been conducted in order to find out which of the eight factors indicates metacognitive skillfulness the best (see Table 4). As the two indicators CVS and the use of an undirected plan had negative directions, the two variables had to be inverted before they could be correlated to the other six measures. This was done by subtracting them from 5. Five was chosen because it is the highest value that could be obtained on the measure of CVS, as not more than 5 (which means all) variables could be changed from one experiment to the next. Therefore, the students changing on average many variables will obtain lower scores on this measure. The same is true for the measure of undirected plans. By subtracting it from five, the students using undirected plans frequently will achieve lower scores. All eight measures were highly correlated. As some of the inquiry learning measured overlap in what they represent, those measures showing the highest correlations for the concept they measure were chosen as indicators for metacognitive skillfulness. The two variables, formulating a hypothesis and formulating a plan were very similar in what they represent to the variables generating a simple hypothesis and checking one variable, respectively. The correlational analysis showed that a combination of four indicators displayed the highest correlation with learning performance. These indicators were: the generation of a simple hypothesis, the use of CVS demonstrating systematic behavior, the plan of checking one variable, and the use of an undirected plan. The correlation between the compound score for metacognitive skillfulness calculated from these measures and learning performance was $r = .32$, $p < .01$.

Significant differences between scores on metacognitive skillfulness of the two age groups were found. The fourth-graders reached a mean score of 5.95. The sixth-graders reached with a mean score of 10.26 significantly higher scores ($F(1, 51) = 13.53$, $p < .01$). No significant differences were found between the scores on the biology and the geography domain, neither for the fourth-graders nor for the sixth-graders.

***Table 4**: Correlational analysis of inquiry learning measures*

		1	2	3	4	5	6	7	8
1	**Formulate Hypothesis**	-	**.93****	.33*	**.71****	**.67****	**.55****	**.72****	.16
2	**Simple Hypothesis**		-	.39**	**.66****	**.68****	**.61****	**.63****	.25
3	**CVS**			-	.44**	**.57****	.22	.49**	**.56****
4	**Formulate plan**				-	**.83****	**.56****	**.90****	**.51****
5	**Checking one variable**					-	.39**	**.93****	**.55****
6	**Check effect of one variable in different experiment**						-	.39**	.27
7	**Inductive plans**							-	.47**
8	**Undirected plans**								-

*p < .05, **p < .01

3.3 Causal mechanisms

In this paragraph, the differences in the generation and the application of causal mechanisms among the two age groups will be discussed. Comparing the means of the number of causal mechanisms generated it became evident, that fourth-graders produced more causal mechanisms than sixth-graders. The difference were significant for the food task ($F (1, 24) = 5.99$, $p < .05$), the total biology domain ($F (1, 50) = 8.97$, $p < .01$) and when both domains evaluated together ($F (1, 104) = 4.22$, $p < .05$). Analyzing both groups individually, no differences in the number of generated causal mechanisms were found for the fourth-graders. The sixth-graders on the other hand, generated significantly more causal mechanisms on the geography domain than on the biology domain.

The number of generated causal mechanisms decreased three out of four times from the first task to the second. However, the difference between the number of causal mechanisms generated on the first and on the second task was only significant for the sixth-grader on the biology domain ($F (1, 22) = 4.53$, $p < .05$).

Table 5: *Means and Standard Deviations of the Presence of Causal Mechanisms on both domains of both age groups*

	Grade 4: CM			**Grade 6: CM**		
	M	**SD**	**n**	**M**	**SD**	**n**
Plant growing task	.34	.27	14	.19	.16	12
Food task	**.36***	.40	14	.07	.12	12
Total Biology	**.35****	.33	14	.13	.15	12
Otter task	.52	.56	14	.43	.34	13
Ageing task	.37	.36	14	.32	.34	13
Total Geography	.45	.47	14	.37	.34	13
Total	**.40***	.41	28	.25	.29	25

Note. Measures are means weighted by the number of experiments conducted. *p< .05, **p< .01

In addition to the analysis of the generation of causal mechanisms, the application of causal mechanisms was investigated. There were no differences found in the application of causal mechanisms between the two age groups ($F (1, 52) = .27$, $p = .61$). Analyzing both domains, the fourth-graders applied on average 0.11 (SD: .14) causal mechanisms per experiment and the sixth-graders 0.10 (SD: 0.13). Evaluating both domains individually, domain-dependent effects were found for the sixth-graders but not for the fourth-graders. On the geography domain, the sixth-graders reached a mean of .15 (SD: .16). On the biology domain however, they obtained a mean of .04 (SD: .06). The mean scores on the geography domain were significantly higher than the scores on the biology domain ($F (1, 24) = 9.61$, $p < .01$). No task-dependent effects were found for the fourth-graders or the sixth graders.

3.4 Relations between causal mechanisms, metacognitive skillfulness, intellectual ability, prior knowledge, causal mechanisms and inquiry learning performance

Pearson's correlations between the generation of causal mechanisms, metacognitive skillfulness, intellectual ability, prior knowledge, and inquiry learning performance across age-groups and domains are presented in Table 6. Intellectual ability, prior knowledge, causal mechanisms, and learning performance were all significantly correlated to metacognitive skillfulness. Its correlation with intellectual ability ($r = .29$, $p < .01$), prior knowledge ($r = .26$, $p < .01$) and learning performance ($r = .32$, $p < .01$) were positive. The

generation of causal mechanisms is the only factor showing a negative correlation with metacognitive skillfulness (r = -.29, p < .01). For intellectual ability, significant correlations were observed with prior knowledge (r = .48, p < .01) and learning performance (r = .35, p < .01). The correlational analysis also revealed a significant positive correlation between prior knowledge and learning performance (r = .27, p < .01). The generation of causal mechanisms showed to be significantly correlated to the application of causal mechanisms (r = .78, p < .01).

***Table 6**: Correlation between metacognitive skillfulness, prior knowledge, intellectual ability, the generation of causal mechanisms and comprehension score across age-groups and domains*

		1	2	3	4	5	6
1	**Metacognitive skillfulness**	-	**.29****	**.26****	**-.29****	-.15	**.32****
2	**Intellectual ability**		-	**.48****	-.01	.05	**.35****
3	**Prior knowledge**			-	.06	.10	**.27****
4	**Generation of causal mechanisms**				-	**.78****	-.07
5	**Application of causal mechanism**					-	.05
6	**Comprehension score**						-

*p < .05, **p < .01

Next, the correlations per age groups will be discussed. The results for the fourth-graders are displayed in Table 7. For them, one significant correlation was found between intellectual ability and prior knowledge (r = .38, p < .01) as well as one between the generation of causal mechanisms and the application of causal mechanism (r = .78, p < .01).

***Table 7:** Correlation between metacognitive skillfulness, prior knowledge, intellectual ability, the generation of causal mechanisms and comprehension score across domains: fourth-graders*

		1	2	3	4	5	6
1	**Metacognitive skillfulness**	-	.03	.00	-.18	-.09	.03
2	**Intellectual ability**		-	**.38****	-.02	-.08	-.08
3	**Prior knowledge**			-	.16	.13	-.09
4	**Generation of causal mechanisms**				-	**.78****	-.11
5	**Application of causal mechanism**					-	.01
6	**Comprehension Score**						-

*p < .05, **p < .01

In Table 8, the results of the correlational analysis for the sixth-graders are presented. Significant correlations were found between metacognitive skillfulness, the generation of causal mechanisms and learning performance. The correlation between metacognitive skillfulness and causal mechanisms was negative ($r = -.34$, $p < .01$), between metacognitive skillfulness and comprehension score positive ($r = .38$, $p < .01$). Intellectual ability was found to be significantly correlated to the generation of causal mechanisms ($r = .32$, $p < .05$), the application of causal mechanisms ($r = .29$, $p < .01$) as well as with comprehension score ($r = .59$, $p < .01$).

Table 8: *Correlation between metacognitive skillfulness, prior knowledge, intellectual ability, the generation of causal mechanisms and comprehension score across domains: sixth-graders*

		1	**2**	**3**	**4**	**5**	**6**
1	**Metacognitive skillfulness**	-	.26	.18	**-.34***	-.20	**.54****
2	**Intellectual ability**		-	.23	**.32***	**.29****	**.42****
3	**Prior knowledge**			-	.22	.16	.18
4	**Generation of causal mechanisms**				-	**.82****	.21
5	**Application of causal mechanism**					-	.16
6	**Comprehension score**						-

*p < .05, **p < .01

Next, the correlations between metacognitive skillfulness, prior knowledge, intellectual ability, the generation of causal mechanisms and comprehension score were analyzed per domain. For the fourth-graders, no variations from the correlation pattern obtained by the analysis of both domains were found. Analyzing the correlations of the fourth-graders of the geography domain, one significant correlation is obtained between the generation of a causal mechanism and the application of causal mechanisms ($r = .88$, $p < .01$), which has been found in all other analyses as well. The analysis of the biology domain did not show any variation from the correlation pattern obtained for both domains, meaning that significant correlations were found between intellectual ability and prior knowledge ($r = .58$, $p < .01$) as well as between the generation and application of causal mechanisms ($r = .54$, $p < .01$).

When the biology and the geography domains were analyzed individually, four of the correlations shown to be significant across domains were also found to be significant for the geography domain. Significant correlations were found between metacognitive skillfulness and the generation of causal mechanisms ($r = -.34$, $p < .05$), metacognitive skillfulness and

comprehension score ($r = .54$, $p < .01$), intellectual ability and the generation of causal mechanisms ($r = .47$, $p < .05$), and between the generation and the application of causal mechanisms ($r = .79$, $p > .01$).

However, the correlation patterns for the sixth-graders on the biology domain showed some variations from the pattern resulting from the analysis of both domains. In addition to the significant correlations between metacognitive skillfulness and comprehension score ($r = .53$, $p < .01$), intellectual ability and prior knowledge ($r = .44$, $p < .05$), and between the generation and the application of causal mechanisms ($r = .73$, $p > .01$), which has also been obtained with the correlational analysis across domains, a significant correlation has been found between the application of causal mechanisms and comprehension score ($r = -.48$, $p < .05$).

Task-dependent effects have only been found for the fourth-graders. On the ageing task, a significant negative correlation was found between metacognitive skillfulness and intellectual ability ($r = -.55$, $p < .05$). On the plant task, a significant correlation was found between prior knowledge and the generation of causal mechanisms ($r = .60$, $p < .05$). Both correlations have neither been obtained by the analyses per domain nor the analysis across domains, indicating task-dependent effects.

The correlation patterns of the individual tasks that have been found for the sixth-graders have also been the obtained by other correlational analyses per domain as well as across domains. Hence, no task-dependent effects exist for the sixth-graders.

4. Conclusion

The first objective of the present study was to learn more about the relationship between inquiry learning behavior and learning performance. First of all, the differences in learning performance were analyzed. It was found that the sixth-graders obtained significantly higher comprehension scores than the fourth-graders, indicating that learning performance increases with age. Evaluating the inquiry learning measures of the total group obtained through the protocol analysis, it is found that formulating hypotheses, formulating simple hypotheses, usage of CVS, making valid inferences, the percentage of valid inferences, formulating plans, checking one variable, checking the effect of a variable in a different experiment, stating inductive plans, and following an undirected plan were significantly correlated to comprehension score. The individual analysis of both age groups showed different correlation patterns. This gives support to our first hypothesis stating that different inquiry measures play a role in the reasoning process of fourth and sixth-graders.

Next, the results of the correlational analysis between inquiry learning measures and comprehension score for the fourth and sixth-graders will be compared to the findings obtained by Wilhelm (2001) with adults. It must be clarified that this comparison is not based on a statistical analysis, but is qualitative in nature. By comparing their correlation patterns, we want to find out if the inquiry learning measures that play a role for children also play a role for adults. Wilhelm used two types of inquiry learning tasks: one was concrete and the other one was abstract. Although the present study also used concrete inquiry learning tasks, the total correlations of the adults on both conditions will be compared to the correlations obtained by the children as they are of greater power than the scores on the individual domains. The correlation pattern obtained for the adults shows great consistency with the correlations obtained for the children. For both children and adults, ten correlations showed to be significant. Eight of them are consistent, two differ from each other. Both age groups reached significant correlations on formulating hypothesis, formulating simple hypothesis, using CVS, making valid inferences, formulating plans, checking the effect of a variable in a different experiment, stating inductive plans, and using an undirected plan. Unlike the children, significant correlations with comprehension score were found for the adults for formulating an inference ($r = .77$, $p < .01$) and for the use of deductive plans ($r = .39$, $p < .01$). As opposed to the adults, the children showed significant correlations between the percentage of valid inferences ($r = .46$, $p < .01$) and checking the effect of one variable ($r = .49$, $p < .01$).

Because of the great consistency of the correlation pattern, it is concluded that the same inquiry learning measures play a role in the reasoning process of children and adults.

In the following, the differences in mean scores on the inquiry learning measures between the fourth and sixth-graders will be analyzed. On almost every observed difference, the sixth-graders scored on average higher. Fourth-graders only reached higher mean scores on the use of undirected plans, indicating a lower level of metacognitive skillfulness than the sixth-graders.

The obtained means on the inquiry learning measures of the fourth- and sixth-graders were compared with the means obtained by the adults (see Table 9). Again, this analysis is qualitative in nature. On 11 out of the 18 measures, the adults reached higher means than the children. Adults and children reached almost the same means on the use of complex hypothesis and the generation of valid inferences. Children reached higher means on formulating inferences and on the generation of a specific outcome. The higher mean score on the formulation of inferences might seem odd at first sight. Evaluation of the outcome is an important aspect of metacognitive skillfulness; therefore adults would be expected to score higher. However, not the mean number of inferences but the percentage of valid inferences indicates high metacognitive skillfulness and adults reach considerately higher mean percentages. Children also varied more variables per experiment, had higher prediction errors and made more use of undirected plans, all indicating lower metacognitive skillfulness. These findings give evidence to the hypothesis that age-dependent effects will be found between the inquiry learning behavior of fourth-graders, sixth-graders and adults. The mean scores on the inquiry learning measures as well as comprehension scores increase with age.

***Table 9**: Means and Standard Deviations of Inquiry Learning Measures: Children and Adults*

	Total Children (n=53) M (SD)	**Total Adults (n=44) M (SD)**
Hypothesis generation		
Formulate hypothesis	.25 (.16)	.45 (.24)
Simple hypothesis	.21 (.14)	.42 (.23)
Complex hypothesis	.04 (.06)	.03 (.05)

Table 9 (continued)

	Total Children (n=53) M (SD)	**Total Adults (n=44) M (SD)**
Design of experiments		
Coverage experiment space	23%	28%
CVS	2.19 (.47)	1.91 (.46)
Valid inference	.63 (.29)	.62 (.31)
Percentage valid inferences	58%	86%
Prediction error	1.02 (.18)	1.01 (.33)
Regulation of Learning		
Formulate plan	.45 (.25)	.75(.20)
Checking one variable	.28 (.20)	.42 (.20)
Checking interaction	.04 (.08)	.06 (.10)
Generate specific outcome	.06 (.11)	.02 (.04)
Check effect of variable in different experiment	.05 (.07)	.17 (.13)
Predict outcome	.03 (.06)	.08 (.09)
Inductive plans	.32 (.23)	.47 (.21)
Deductive plans	.12 (.13)	.25 (.16)
Undirected plans	.55 (.25)	.26 (.19)

The second objective of the study was to investigate the relations between causal mechanisms, metacognitive skillfulness, intellectual ability, and prior knowledge with learning performance as well as how they relate to each other.

First of all, the findings on causal mechanisms will be presented. Fourth-graders generated significantly more causal mechanisms than sixth-graders confirming the hypothesis that the use of causal mechanisms will decrease with age. The generation of causal mechanisms dropped from the first to the second task with the exception of the fourth-graders in the biology condition. However, the decrease was only significant for the sixth-graders in the geography condition. The mean application of causal mechanisms to the task showed no significant difference between the two age groups. A significant positive correlation was found between the generation and the application of causal mechanisms for

the total group as well as for the fourth and sixth-graders, when evaluated individually. From that we can conclude that if causal mechanisms are generated, they are also likely to be applied in the following experimentation process. To represent the inquiry learning measures by a single variable indicative of metacognitive skillfulness, four of the significantly correlating inquiry learning measures have been combined. Evaluating both age groups together, the generation of causal mechanisms showed a significant negative correlation with metacognitve skillfulness. Our hypothesis stating that the generation of causal mechanisms is an indicator for low levels of metacognitive skillfulness is therefore confirmed. The correlational analysis of the plant task for the fourth-graders found a significant positive relationship between prior knowledge and the generation of causal mechanisms. The analysis of the biology domain for the sixth-graders displayed a significant negative correlation between the application of causal mechanisms and comprehension score. Unexpectedly, a significant positive correlation was found between intellectual ability and the generation and the application of causal mechanisms for the sixth-graders across domains.

The correlational analysis between metacognitive skillfulness and comprehension score was found to be significant for the total group and the sixth-graders, but not for the fourth-graders. Therefore, it can be assumed that the significant correlation obtained for the total group is due to the high correlation of the sixth-graders. From that we can conclude that the effect of metacognitive skillfulness on learning performance emerges with age. When both age groups are evaluated together, metacognitive skillfulness displays a significant positive correlation with intellectual ability. However, the correlational analysis of the ageing task of the fourth-graders found a negative correlation between the two variables. From the obtained results it can be concluded that the relationship between metacognitive skills and intellectual ability depends on age. Our hypothesis that a positive correlation will be found between intellectual ability and metacognitive skillfulness is therefore only confirmed for the older students. No hypothesis was stated about the correlation between prior knowledge and metacognitive skillfulness as it was exploratory. Metacognitive skillfulness showed to be significantly correlated to prior knowledge for the total group. Consequently, it can be said that learners with high levels of metacognitive skillfulness also tend to have high levels of prior knowledge. However, when analyzed individually, neither the fourth nor the sixth-graders displayed a significant correlation between the two variables.

Next, the effects of intellectual ability will be presented. As indicated by the significant correlation between intellectual ability and comprehension score, higher levels of intellectual ability have a positive effect on learning performance. The correlation between intellectual

ability and comprehensions score was significant for the total group and the sixth-graders, but not for the fourth-graders. Again, it can be assumed that the significant correlation obtained for the total group is due to the high correlation obtained by the sixth-graders. Consequently, the hypothesis that intellectual ability affects learning performance positively is only confirmed for the older students. In other words, the influence of intellectual ability on learning performance depends on age. A significant correlation was found for the between intellectual ability and prior knowledge for the total group, as well as for the fourth-graders. As the sixth-graders also displayed a positive although not significant correlation between the two variables ($r = .23$) the significant correlation found for the total group is not ascribed to the fourth-graders. Accordingly, we conclude that, independent of their age, learners with high levels of intellectual ability also possess high levels of prior knowledge.

A significant positive correlation between prior knowledge and comprehension score has been obtained for both age groups as well as for the sixth-graders. Once again, we can act on the assumption that the significant correlation found between prior knowledge and comprehension score for the total group results from the significant correlation found for the sixth-graders. Accordingly, it can be concluded that the way prior knowledge and learning performance correlate with each other is age-dependent.

5. Discussion

There is some disagreement on which factors belong to the concept of metacogntitive skillfulness (Schneider, 2009). It can be argued that CVS might not be a valid indicator for metacognitive skillfulness as it is also part of the concept of metacognitive knowledge as it can be learned. A study has shown that children who received a lesson on the use of CVS outperformed students who did not receive the lesson (Sao Pedro, Gobert, Hefernan & Beck, 2009). Hence, the knowledge about CVS resulted in the use of the CVS. In the present study it was not assessed whether the learners knew about the CVS but if they actually applied the strategy to the task. Therefore, it can be argued that the aspect of CVS actually represented metacognitive skillfulness and not metacognitive knowledge, as only the pure application of the strategy to the presented task has been assessed.

As stated before, the generation of causal mechanisms plays an important role for evaluating correlations present in daily life (Koslowski, 1996). However, research has shown that in inquiry learning the generation of causal mechanisms plays only a minor role (Kuhn, 1989). Koslowski (1996) argues that the reason for the low use of causal mechanisms is the design of the inquiry learning tasks. The design does not foster the learner to reason in a broad way about possible theories and considering evidence from covariation as well as from causal mechanisms. In the present study, the use of causal mechanisms was quite modest. A reason for this result could be the involvement of children. As said, younger children tend to generate more causal mechanism than older ones (Koslowski, 1996). Earlier results are supported by the present study finding that fourth-graders generated significantly more causal mechanisms than the sixth-graders. When evaluated across domains, the fourth-graders generated in 40% of the experiments causal mechanisms, as opposed to 25% for the sixth-graders. It was also found that the sixth-graders assigned to the geography domain used significantly more causal mechanisms than the sixth-graders assigned to the biology domain. Before the start of the task, the students were asked to state their expectations about how each variable would influence the outcome. Students conducting the tasks of the geography domain seemed to have more trouble making prediction about the effects of the variables. Koslowski (1996) argues that, although a person has no knowledge about a mechanism that could explain an observed covariation, it is assumed that a causal mechanism underlies the covariation that has just not been discovered yet. Therefore, it may be assumed that the students tried to make sense of the variables by stating causal mechanisms that could offer an explanation for the observed outcome.

Moreover, it was found that the number of generated causal mechanisms decreased almost every time from the first to the second task the students conducted (with the exception of the biology tasks carried out by the fourth-graders where numbers stayed almost constant). However, the decrease of causal mechanisms was only significant for the sixth-graders on the geography domain. This indicates that only the older students became test-wise in the course of the experiment, more and more disregarding the presence of causal mechanisms that would offer an explanation for the obtained outcomes.

Using causal mechanisms as evidence in inquiry learning tasks is often regarded as flawed scientific reasoning (Kuhn, 1989). This study has shown that the use of causal mechanisms is negatively related to metacognitive skilfulness for the older students. Metacognitive skillfulness has been found to be a good predictor of learning performance in earlier studies (e.g. Veenman, Wilhelm & Beishuizen, 2004) as well as in the present study, as metacognitive skillfulness was significantly correlated with comprehension score for both domains and the sixth-graders. These two observations would support the assumption that the use of causal mechanisms somehow suppresses metacognitive skillfulness which in turn leads to a weaker learning performance. This was only found to be true for the sixth-graders, but not confirmed for the fourth-graders. The generation of causal mechanisms does neither weaken their learning performance nor indicates low metacognitive skillfulness. The role of causal mechanisms and metacgonitive skillfulness therefore depends on age. It needs to be stressed, that the conclusions drawn only refer to inquiry learning. No inference can be made on the effect and usefulness of causal mechanisms in the real world or in different experimental settings.

As said, causal mechanisms play a greater role for younger students. As causal mechanisms are based on prior knowledge and the sixth-graders in this study possess significantly more prior knowledge about the two domains, it would be expected that they also generate more causal mechanisms. However, the results of the present study support the findings obtained by Kuhn et al. (1988) which has shown that older students used significantly less causal mechanisms in inquiry learning tasks. As stated above, older students are possibly more test-wise as they realize that the prior domain knowledge they possess does not play a role for the successful conduction of inquiry learning tasks. The finding that the sixth-graders reach higher scores on learning performance gives evidence that not suspending but using prior knowledge thoughtfully is beneficial.

Furthermore, it has been found that the use of causal mechanisms is negatively correlated to metacognitive skillfulness. This means that students who generate more causal

mechanisms, score lower on metacognitive skillfulness. A significant negative correlation was not found for the fourth-graders implying an age-dependent effect. Fourth-graders have shown to score considerably lower on metacognitive skillfulness as well as higher on the use of causal mechanism. It can be concluded that, when the level of metacognitive skillfulness is low, causal mechanisms are used more frequently and do not affect learning performance negatively. In children with higher metacognitive skills, the generation of causal mechanisms hinders good scientific reasoning. Even though causal mechanisms are based on prior knowledge which was found to enhance learning performance in the present study, the presence of causal mechanisms hinders learning performance via metacognitive skillfulness.

It was found that intellectual ability only has a positive effect on comprehension score for the sixth-graders but not for the fourth-graders. An explanation for the absence of a significant correlation between intellectual ability and comprehension score is the threshold of problematicity theory (Elshout, 1987). This theory implies that the influence of intellectual ability on learning performance depends on the complexity of the task. When the complexity of a task is very high or very low, intellectual ability tends to show no effect on learning performance at all. In the present study, the fourth-graders obtained significantly lower scores on comprehension than the sixth-graders ($F(1, 52) = 18.30$, $p < .01$), although the level of difficulty was aimed to be the same for both age groups. Therefore, the complexity of the task could have been too high for the fourth-graders, explaining the absence of a significant relationship between intellectual ability and learning performance. The significant correlation between intellectual ability and comprehension score found for the sixth-graders implies that the level of complexity of the task was neither too high nor to low, indicating an impact of intellectual ability on their learning performance.

Unexpectedly, intellectual ability showed to be significantly correlated to the generation and the application of causal mechanisms for the sixth-graders. This is especially interesting as intellectual ability shows to be positively correlated to metacognitive skillfulness, which in turn is significantly negatively correlated to the generation of causal mechanisms. Moreover, a significantly negative correlation was found for the sixth-graders on the biology domain between the application of causal mechanisms and comprehension score. The application of causal mechanisms can therefore be regarded as an indicator of poor learning performance. These contradicting findings are very difficult to explain given the small amount of research that has been conducted on this issue. Further research is needed to gain more insight into the relationships between the different concepts.

Next, the findings on prior knowledge will be discussed. The correlation between prior knowledge and learning performance was found to be significantly positive for sixth-graders. Research has shown that high levels of prior domain knowledge have a positive effect on learning performance (Lazonder, Wilhelm, & Van Lieburg, 2009). This finding is confirmed by the present study. Sixth-graders reached significantly higher scores on the prior knowledge test ($F(1, 52) = 26{,}359$, $p < .01$). So, not only the level but also the use of prior knowledge depends on age. The present study implies that sixth-graders make better use of their prior knowledge, possibly in the generation of relevant hypothesis and theories about how the variables relate to each other. As research has shown, the coordination of theory and evidence is more difficult for younger children. Fourth-graders often have trouble distinguishing between the theories they hold and the evidence provided (Koslowski, 1996). Consequently, they face difficulties in disregarding their initial theories although the obtained evidence speaks against them. Therefore, not only the lower level of prior knowledge but also the lack of skills on how to use prior knowledge effectively, affects the learning performance of the fourth-graders negatively. A significant correlation between prior knowledge and the generation of causal mechanisms was only obtained for the plant task conducted by the fourth-graders. In this task, higher levels of prior knowledge indicate a frequent use of causal mechanisms. The finding is noteworthy because, as it has been mentioned before, the causal mechanisms one generates are often based on the prior knowledge the learner holds.

The differences in the children's abilities in giving information about their reasoning process could also partly be responsible for the obtained results. It is possible that the younger students had to invest more cognitive effort to the requirement of thinking aloud which in turn leaves less cognitive effort to demand to the actual task. In addition, being able to express one's path of thought soundly might result in higher comprehension scores as more of the learners reasoning process could be captured and evaluated. For that reason, high comprehension scores might not always be evident for high cognitive but for verbal skills. It might be interesting to include a test of verbal predisposition in further experiments, in order to investigate its influence on inquiry learning tasks.

Next, the experimental setup of the present study will be discussed. Differences within each of the two age groups could be controlled by the prior assessment of domain knowledge and intellectual ability. The students were randomly assigned to the two conditions which also heightened the internal validity of the study. However, the significant difference in comprehension score between the two domains that has been found for the fourth-graders

implies that the level of difficulty of the two domains is not the same. Moreover, several task-dependent effects have been observed, although it must be noted that the power of the results obtained per task is generally low due to the small sample sizes conducting each task. As the level of prior knowledge and intellectual ability was shown to be equal among the age groups, differences have to exist between the contents of the tasks. These differences seem to lead to the use of different approaches and strategies in the conduction of the tasks.

The interrater reliability between the evaluations of the thinking-aloud protocols were quite high. However, it must be noted that only 5% of the protocols were assessed by two raters. Moreover, both raters knew about the aim of the study. The reliability of the scoring could have been increased trough the use of a third independent rater who did not know about the objectives of the study. This would have lowered the impact of evaluation bias.

The external validity is lowered by the fact that the students who participated in the study all go to the same school, live in the same city, in the same country, and probably have the same social background. The replication of this study with a different social group might lead to different results.

Another aspect that could lower the validity of the obtained results is the experimenter bias. Research by e.g. Rosenthal and Jacobson (1992) has shown that especially young children are very sensitive for the expectation the experimenter has of them. In the original analysis, the factor causal mechanism was not measured. When causal mechanisms were generated during the tasks, the experimenter did not rate or respond to it in any possible way that could have shown his expectancy about the generation of causal mechanisms to the students. As the use of causal mechanisms was not expected in the original study, the experimenter's expectancy could not have influenced the students in this regard. However, the experimenter's expectancy could have had an impact on the scientific reasoning process of the students through unconscious signals. In the present study, the experimenter used verbal encouragements and praise to keep the learner motivated. Although encouragements were given to learners showing good scientific reasoning as well as to learners who did not, the learners might nevertheless recognized whether they deserve the praise they received or not. As a result, their learning behaviour might have been influenced.

As said in the beginning, inquiry learning offers students the opportunity to construct knowledge themselves. This raises the question: What did the students learn from the conduction of these inquiry learning tasks? Although the model of the tasks corresponded to possible real-life phenomena, it is doubted that the students will remember the particular effects the independent variables had on the dependent variable. Boyd (cited by Koslowski,

1996) stated that the evaluation of covariation without applying prior knowledge through the generation of causal mechanisms did often result in low learning outcomes. When no prior knowledge is taken into account while conducting inquiry learning tasks, the task content does not become meaningful to the subject. As a result, the subject does not remember the outcomes of the task. However, although the students might not have obtained domain knowledge through the inquiry learning tasks, they might have increased their metacognitive skillfulness.

In the present study a correlational analysis was used to learn more about the relationships between metacognitive skillfulness, prior knowledge, intellectual ability, causal mechanisms and learning performance. This statistical method is restricted. The use of a regression analysis could shed some more light on how the concepts relate to each other and how the relationships change when one of the variables is varied. Further research should therefore consider the use of a regression analysis.

References

Ahn, W., & Kalish, C. W. (2000). The role of mechanism beliefs in causal reasoning. In F. C. Keil & R. A. Wilson (Eds.), *Cognition and explanation (pp 199-225)*. Cambridge, MA: MIT Press.

Anderson, S. (2002). Working together to develop a professional learning community. In *Proceedings of the 2002 HERDSA Conference (pp. 20–27)*, Perth, Australia. Higher Education Research and Development Society of Australia.

Busato, V. V., Prins, F. J., Elshout, J. J., & Hamaker, C. (2000). Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education. *Personality and Individual Differences, 29*, 1057–1068.

Chen, Z., & Klahr, D. (1999). All other things being equal: acquisition and transfer of the control of variables strategy. *Child Development, 70*, 1098-1120.

De Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research, 68*, 179–201.

Eindtoets Basisonderwijs 1998. CITO, Arnhem 1997. Flanagan, J. C. (1951-1956). *Flanagan Aptitude Classification Tests.* S.R.A.

Elshout, J. J. (1976). *Karakteristieke moeilijkheden in het denken.* Unpublished doctoral dissertation, Universiteit van Amsterdam.

Elshout, J. J. (1983). Is measuring intelligence still useful?. In S. B. Anderson, & J. S. Helmick (Eds.), *On educational testing (pp. 45–56)*. San Francisco: Jossey-Bass.

Elshout, J. J. (1987). Problem solving and education. In E. De Corte, H. Lodewijks, R. Parmentier, & P. Span (Eds.), *Learning and instruction (pp. 259-273)*. Oxford, UK/Leuven, Belgium: Pergamon Books/University Press.

Evers, A., & Lucassen, W. (1983). *Differentiële Aanleg Testserie.* Lisse, The Netherlands: Swets & Zeitlinger.

Examen Biologie VBO-MAVO-C 1996-1997. CITO, Arnhem.

Examen Biologie VBO-MAVO-D 1996-1997. CITO, Arnhem.

Examen Aardrijkskunde VBO-MAVO-C 1996-1997. CITO, Arnhem.

Examen Aardrijkskunde VBO-MAVO-D 1996-1997. CITO, Arnhem.

Hulshof, C. D. (2001). Discovery of ideas and ideas about discovery: *The influence of prior knowledge on scientific discovery learning in computer-based simulations.* Unpublished doctoral dissertation, University of Twente, The Netherlands.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*, 1-48.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning.* Cambridge, MA: MIT Press.

Koslowski, B., & Okagaki, L. (1986). Non-Humean indices of causation in problem-solving situations: Causal mechanisms, analogous effects, and the status of rival alternative accounts. *Child Development, 57,* 1100–1108.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological review, 96*, 674-689.

Kuhn, D., Garcia-Mila, M., Zohar, A., & Andersen, C. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development, Serial No. 245, 60* (40), 1–128.

Kuhn, D., Katz, J., & Dean, D. (2004). Developing reason. *Thinking & Reasoning, 10*, 197 – 219.

Kuhn, D., & Pease, M. (2008). What needs to develop in the development of inquiry skills? *Cognition and Instruction, 26*, 1–48.

Lazonder, A. W., Wilhelm, P., & van Lieburg, E. (2009). Unraveling the influence of domain knowledge during simulation-based inquiry learning. *Instructional Science,* 37, 437-451.

Löhner, S., van Joolingen, W. R., & Savelsbergh, E. R. (2003). The effect of external representation on constructing computer models of complex phenomena. *Instructional Science, 31*(6), 395–418.

Mommers, J. C., Visser, Chr. de, & Warnars, R. L. J. (1977). *Uitkomst. Rekenmethode door een werkgroep onder leiding van Cas Klavier. 6B. Repetitieboek: boek voor de onderwijzer.* Tilburg: Uitgeverij Zwijssen B.V.

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.

Prins, F. J., Veenman, M. V. J., & Elshout, J. J. (2006). The impact of intellectual ability and metacognition on learning: New support for the threshold of problematicity theory. *Learning and Instruction, 16*, 374– 387.

Rosenthal, R., & Jacobson, L. (1992). *Pygmalion in the classroom* (expanded edition). New York: Irvington.

Saab, N., Van Joolingen, W. R., & Van Hout-Wolters, B. H. A. M. (2005). Communication in collaborative discovery learning. *British Journal of Educational Psychology, 75*, 603-621.

Sao Pedro, M., Gobert, J., Heffernan, N. & Beck, J. (2009). *Comparing Pedagogical Approaches for Teaching the Control of Variables Strategy.* Cognitive Science Society Annual 2009 Conference.

Schneider, W. (2009). Metacognition and memory development in childhood and adolescence. In H.S. Walters & W. Schneider (Eds.), *Metacognition, strategy use, and instruction* (pp. 54-81). New York: Guilford Press.

Van Joolingen, W. R., de Jong, T., & Dimitracopoulou, A. (2007). Issues in computer supported inquiry learning in science. *Journal of Computer Assisted Learning, 23*(2), 111-119.

Van Joolingen W.R., de Jong T., Lazonder A.W., Savelsbergh E.R. & Manlove S. (2005) Co-Laboratory: research and development of an online learning environment for collaborative scientific discovery learning. *Computers in Human Behavior, 21*, 671–688.

Van der Stel, M. & Veenman, M.V.J. (2007). Relation between intellectual ability and metacognitive skilfulness as predictors of learning performance of young students performing tasks in different domains. *Learning and Individual Differences, 18*, 128–134

Veenman, M. V. J., & Beishuizen, J. J. (2004). Intellectual and metacognitive skills of novices while studying texts under conditions of text difficulty and time constraint. *Learning and Instruction, 14*, 621−640.

Veenman, M. V. J., Beishuizen, J. J., & Niewold, P. (1997). Intellectual ability and metacognitive skills during text studying. In S. Vosniadou, E. Matsagouras, K. Maridaki-Kassotaki, & S. Kotsanis (Eds.), *Abstracts of the 7th European Conference for Research on Learning and Instruction (pp. 145–146)*. Athens: Gutenberg UP.

Veenman, M. V. J., Prins, F. J., & Elshout, J. J. (2002). Initial inductive learning in a complex computer simulated environment: the role of metacognitive skills and intellectual ability. *Computers in Human Behavior, 18*, 327-342.

Veenman, M. V. J., & Spaans, M. A. (2005). Relation between intellectual and metacognitive skills: Age and task differences. *Learning and Individual Differences, 15*, 159−176.

Veenman, M.V.J., Wilhelm, P. and Beishuizen, J.J. (2004). 'The relation between intellectual and metacognitive skills from a developmental perspective', *Learning and Instruction, 14*, 89–109.

Wilhelm, P. (2001). Knowledge, skills and strategies in self-directed inductive learning. Doctoral dissertation, Leiden University, The Netherlands.

White, B. Y., & Shimoda, T. A. (1999). Enabling students to construct theories of collaborative inquiry and reflective learning: computer support for metacognitive development. *International Journal of Artificial Intelligence in Education, 10*, 151–182.

APPENDIX

Appendix A: Scoring Protocol

- **De naam van de variabele in SPSS staat tussen haakjes**

A. Proefpersoonnummer (ppn)

B. Sexe

Man: 1
Vrouw: 2

C. Groep

Groep 6: 6
Groep 8: 8

D. Conditie

Aardrijkskunde - Biologie: 1
Biologie - Aardrijkskunde: 2

E. Starttijd nieuw experiment (startEXP)

Definitie: een experiment begint en eindigt op het moment dat een proefpersoon een niveau van de eerste variabele in een nieuw experiment aanklikt.

- voer het tijdstip (in minuten en seconden) in waarop bovenstaande gebeurt.
- meestal (maar niet altijd) is de vraag "Wat ga je nu uitzoeken?" het begin van een nieuw experiment

F. Experiment (experiment)

Voer het volgnummer in.

G. Plan (plan)

Scoor 1 van de volgende 6 categorieën:

Algemeen: overtuiging is belangrijk
1. Onderzoek het effect van een variabele
("Ik ga uitzoeken wat de fietsen uitmaken")
De variabele waar het om draait ("focus") moet hierbij ook gescoord worden (categorie H.).
2. Onderzoek het gezamenlijk effect van twee variabelen.
("Ik ga uitzoeken wat de fietsen te maken hebben met de schoenen")
De variabelen waar het om draait ("focus") moeten hierbij ook gescoord worden (categorie H.).

- Plan 1 of 2 worden gescoord op basis van het antwoord op de vraag: "Wat ga je nu uitzoeken?" Als in het voorafgaand experiment gezegd wordt wat er nog moet worden uitgezocht en een ppn. gaat dat (op basis van de logfiles) ook doen, dan scoor je ook plan 1 of 2.

- Als iemand netjes aan het experimenteren is (bv. als door de keuze van de variabelen duidelijk wordt welke plan er wordt gevolgd ook als de variabele niet expliciet wordt genoemd) wordt eerder plan 1 of 2 gescoord dan plan 6. Echter, de log-files dienen alleen ter ondersteuning van de protocollen. Enkel en alleen van de log-files kan niet afgeleid worden of er een plan en welk plan er wordt gevolgd.

3. Een specifieke uitkomst genereren (= doel).
   Middel: geleerde regels toepassen.
   (als het plan is om een specifieke uitkomst te genereren, dan moet die ook genoemd worden. Er mag niet alleen gezegd worden dat er een heel goede of slechte uitkomst gegenereerd gaat worden).
4. Verificatie - I
   Doel: kijken of het effect van een bepaalde variabele hetzelfde is in een ander experiment (regel t.a.v. één variabele verifiëren). Middel: nieuw experiment.

Deze categorie heeft een focus en zeer waarschijnlijk ook een hypothese.

5. Verificatie - II
   Doel: de uitkomst van een rijtje te voorspellen op basis van wat je weet (regels t.a.v. alle variabelen toetsen). Middel: nieuw experiment
   Deze categorie heeft geen focus!
6. Overig
   - Informatie verzamelen
     Doel: nieuwe informatie verzamelen, bijv. omdat iemand is vastgelopen
     Voorbeeld: "Kijken wat er hier uitkomt."
   - Als de plan-vraag niet is gesteld of uit de experimenten niet duidelijk is wat pp. van plan is.
   - Idiosyncratisch plan (plan wat niet te scoren is in 1 - 5)

H. Focus plan: welke variabelen zijn bij het plan betrokken? (focusVAR)

(geldt alleen als plan = 1, 2 of 4, anders focus plan = 99).

Focus variabele x: x (1,2,3,4,5)
Focus variabele x en y: xy
Geen focus (of n.v.t.): 99

X en Y verwijzen naar de nummers die onder punt T. worden genoemd!

I. Hypothesen stellen (hypothese)

Ten aanzien van variabele x: x
Ten aanzien van variabele x en y: xy
Geen hypothese (of n.v.t.): 99

Wanneer is een uitspraak een hypothese?
Algemeen: Overtuiging en duidelijkheid zijn belangrijk
Criterium: er moet worden verwezen naar specifieke variabelen en aangegeven worden dat (minimaal 1 van onderstaande situaties):

- een variabele een verschil maakt.
- een variabele geen verschil maakt.
- het effect van een variabele een bepaalde richting heeft.
- het effect een bepaalde grootte heeft.

De uitspraak: "als dit experiment een andere uitkomst oplevert dan maakt die variabele verschil", merk je niet aan als hypothese.

Niet scoren: hypothesen ten aanzien van meer dan twee variabelen (= 99)

X en Y verwijzen naar de nummers die onder punt T. worden genoemd!

J. Inferentie: conclusies (inferentie)

Algemeen: overtuiging is belangrijk

Inclusie inferentie:	1
Exclusie inferentie:	2 (t.a.v. variabele 3: ook scoren)
Interactie:	3 (de uitspraak: “Er is geen interactie tussen…” ook zo scoren)
Geen inferentie	0

- Er moet een verwijzing naar een variabele zijn (dus niet enkel een niveau) of een indicatie dat er sprake is van twee niveaus (bijv. door gebruik vergrotende trap, “verschil” of “bepalend”). Als er bijv. wordt gezegd “Meerdere gebieden zijn beter” is dat al een inferentie.
- Scoor alleen duidelijke/overtuigde inferenties, die volgen op de inferentie-vraag. Scoor samenvattingen van eerdere conclusies (bijv. bij het maken van aantekeningen) niet, tenzij een eerdere conclusie wordt veranderd of een nieuwe inferentie wordt gegeven voor de (nieuwe) inferentie-vraag is gesteld. Wanneer iemand netjes experimenteert scoor je sneller een inferentie.
  Een interactie effect gaat voor een hoofdeffect op één van de betrokken variabelen (eerdere conclusie wordt dan a.h.w. veranderd).
- Wanneer een eerdere inferentie correct is maar niet valide en daarna volgt dezelfde inferentie maar dan valide, dan scoor je deze allebei.
- Wanneer iemand zegt: “Ik heb uitgevonden dat de uitkomsten verschillend zijn”, of “Deze (dit experiment) is minder dan die”, dan scoor je bij inferentie: 0. Het gaat nl. alleen maar om inferenties ten aanzien van één of meerdere variabelen. Als er geen variabelen worden genoemd is er dus geen inferentie gemaakt.
- Wanneer iemand zegt: “X en Y variabele zijn bepalende factoren”, dan scoor je een inferentie. Dit is een inclusie-inferentie

Wanneer een pp. bij een experiment meerdere inferenties maakt moeten deze inferenties in een bepaalde volgorde worden ingevoerd, nl. de variabele met de laagste focuswaarde eerst. Bij een interactie staat het kleinste getal altijd vooraan in de focus (dus altijd 12 en nooit 21 schrijven). Het eerste en het tweede getal zijn bepalend voor de volgorde waarin de inferenties worden opgetekend in de datafile.

Voorbeeld:	1	1	1	
	12	2	4	
	23	3	5	
	3	4		etc.

Bij twee inferenties t.a.v een variabele (komt voor bij variabele 3) bepaalt de code van het type inferentie de volgorde in de datafile. Dat betekent dat een inclusie-inferentie (1) eerst wordt ingevoerd en daarna de exclusie-inferentie (2).

Wanneer iemand een inclusie/exclusie inferentie maakte en daarbij zei: “in deze combinatie”, “met de rondjes” etc. heb ik niet anders gescoord. Deze toevoegingen zijn conservatief en dus prima.

K. Focus inferentie (effect codes) (focusINF)

Focus variabele x: x
Focus variabele x en y: xy
N.v.t.: 99

Criterium: wordt alleen gescoord als Plan 1, 2 of 4 gevolgd worden.

X en Y verwijzen naar de nummers die onder punt T. worden genoemd!

L. Valid: gegeven verzamelde evidentie (validINF)

Indien valide:	1
Indien invalide:	0
N.v.t. (als er geen inferentie werd gemaakt)	99

Invalide zijn die inferenties die niet stroken met gegenereerde evidentie. Meestal betekent dit dat geen rekening wordt gehouden met het constant houden van variabelen (veronderstelde irrelevante effecten worden geacht ook niet te interacteren). Soms wordt een eerdere (valide) inferentie veranderd. Dit gebeurt voornamelijk bij de interacterende variabelen. Een zodanig inferentie is invalide als er elders het (valide) tegendeel ligt. Om de validiteit te beoordelen moet naar alle eerdere experimenten gekeken worden.

M. Correctness (t.a.v. het model in de taak) (correctINF)

Correct:	1
Incorrect:	0
N.v.t.:	99

Bij deze categorie staat vaak een “00” bij valide inclusie/exclusie inferenties ten aanzien van de interacterende variabelen. Die vind ik niet zo interessant, het geeft alleen maar aan dat pp. op dat moment op een dwaalspoor kunnen zitten. Met name interessant zijn de interactie - inferenties (12) die valide zijn en correct volgens het model. Als iemand een interactie onderzoekt en concludeert: er is geen interactie tussen variabele 4 en 5 (klopt) en dat valide weet af te leiden, dan scoor je een interactie - inferentie die valide is en correct volgens het model.

N. Causaal Mechanisme (CM)

Aanwezig:	1
Afwezig:	0

<u>Algemeen:</u> overtuiging is belangrijk.
<u>Definitie:</u> een causaal mechanisme ondersteunt een hypothese en biedt een verklaring voor een verband tussen variabelen. Er moet verwezen worden naar een specifieke hypothese over een of twee variabelen.

- Bij het genereren van een causaal mechanisme wordt aanvullende informatie gegeven om de relatie tussen twee variabelen te verklaren (bijv. kleine pot en veel water is slecht omdat de plant dan verdrinkt).
- Het causale mechanisme kan aangepast worden aan de uitkomsten van de experimenten.
- Het causale mechanisme kan spontaan of als reactie op een prompt gegenereerd worden.

O. Frequentie Causaal Mechanisme (CMfrequentie)

Aantal CM:	1, 2, 3, …
Nvt.:	0

Het gaat hier om het aantal gegenereerde causale mechanismen per experiment.

- Als er geen causaal mechanisme wordt gegenereerd, scoor je een “0”.
- De causale mechanismen worden per experiment gescoord (zie definitie punt E)

P. Uniekheid van het causaal mechanisme (CMrang)

Niet uniek: 0
Uniek: rangnummer geven (1, 2, 3, ….)

Elk nieuw causaal mechanisme (in relatie tot alle vorige) krijgt een nieuw rangnummer.

- Als een unieke causaal mechanisme wordt gegenereerd wordt een rangnummer gegeven, altijd een punt hoger dan de laatste waarde in de SPSS-kolom
- Als er meerdere (unieke) causale mechanisme worden gegenereerd in een experiment wordt dat aantal opgeteld bij het hoogste rangnummer dat in de SPSS-kolom staat. Het resulterende aantal wordt ingevoerd bij het experiment.
- Als een causaal mechanisme niet meer uniek is wordt het genegeerd

Q. Evidentie betrekken bij het opstellen van een theorie (hypothese + causaal mechanisme)

Betrekken: 1
Niet betrekken: 0
Nvt: 99

Een causaal mechanisme kan het experimenteergedrag beïnvloeden.

Criterium: het experimenteergedrag in het eerstvolgende experiment of daarna wordt aangepast op basis van een gegenereerd causaal mechanisme (deductief proces).

- Het is niet voldoende een causaal mechanisme te genereren op basis op het resultaat van het rijtje warmee de proefpersoon bezig is. Dit wordt al door punt N uitgedrukt.
- Als er geen causaal mechanisme wordt gegenereerd is het niet van toepassing en er wordt 99 gescoord.

Appendix B: Plant task

Please read the following instructions carefully.

Plant growing

Plants usually grow from being very small to being fairly big. However, not all plants of the same kind become equally big. How is this possible?
How much water is given might make a difference, for example once or twice a week.
The usage of an insecticide to keep away plant louses and putting some dead leaves in the flower pot also make a difference. The place were the plant grows could also be important. You can put the plant in a greenhouse, inside a house, or on a balcony. The size of the flowerpot might also make a difference. You can put the plant in a big or a small flowerpot.
As you can see, there are several choices you can make when you let a plant grow. The purpose of this task is to find out how these choices influence the growth of the plant.
The plant in this task can grow to become 5, 10, 15, 20 or 25 CMS long.

Appendix C: Model Plant task

Model plant-growing task

	Small flower pot Water once/wk	Small flower pot Water twice/wk	Big flower pot Water once/wk	Big flower pot Water twice/wk
Inside house	20	5	10	10
Greenhouse	25	10	15	15
Balcony	25	10	15	15

Note. once/wk: once per week, twice/wk: twice per week. Usage of an insecticide and putting dead leaves in the flowerpot represent irrelevant variables and are not included in this scheme. Figures represent CMS.