COMPUTERS & EDUCATION

Computers & Education 42 (2004) 389–402

www.elsevier.com/locate/compedu

Computer-supported inquiry learning: effects of training and practice

Jos Beishuizen[a,*], Pascal Wilhelm[b], Marieke Schimmel[c]

[a] Vrije Universiteit Amsterdam, Centre for Educational Training, Assessment and Research, De Boelelaan 1105,
1081 HV Amsterdam, The Netherlands
[b] University of Twente, The Netherlands
[c] Leiden University, The Netherlands

Received 8 January 2003; accepted 29 September 2003

Abstract

Inquiry learning requires the ability to understand that theory and evidence have to be distinguished and co-ordinated. Moreover, learners have to be able to control two or more independent variables when formulating hypotheses, designing experiments and interpreting outcomes. Can sixth-grade (9–10 years) children be trained to acquire these inquiry learning skills? Or is the opportunity to practice in a computer-supported simulation environment a sufficient condition to foster inquiry learning skills? In this study, two groups of sixth grade children were compared: a training group, and a practice group. The training group received an off-line inquiry learning training in which we focused on fostering strategies for proper inferencing and designing experiments. The practice group conducted four inquiry learning tasks during two practice sessions. Learning outcomes and inquiry learning process measures were collected to study whether training and practice resulted in desired changes in learning behaviour. Both training and practice resulted in better performance during the test problems. Compared to the practice group, the training group showed some advantage in discovering an interaction effect. Practice and training effects appeared to be dependent on type of domain.

Inquiry learning is the process of discovering rules governing the relations between independent and dependent variables on the basis of experiments in which the independent variables are manipulated (Wilhelm, 2001). For instance, Chen and Klahr (1999) asked children of 9 and 10 years of age to figure out how far strings stretch by comparing strings with different length, coil width and wire diameter. The children compared the strings two by two by hanging them on hooks on a frame, selecting weights, hanging the weights on the springs and observing as the springs stretched. Computer-based simulation environments offer excellent opportunities for

---

* Corresponding author. Tel.: +31-20-4449201.
E-mail address: j.beishuizen@ond.vu.nl (J.J. Beishuizen).

0360-1315/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compedu.2003.10.003

inquiry learning (De Jong & Van Joolingen, 1998; De Jong et al., in press). Students are allowed to explore a virtual world, manipulate variables, observe the effects of their operations, and conduct experiments to discover relations between variables. During the tasks used in this experiment participants investigated the influence of environmental factors on the growth of a plant, they studied the way eating and drinking habits affect a person's health, they explored the influence of environmental factors on the health of otters, and they studied factors affecting the age of a population in a certain country. (Kuhn 1989; Kuhn, Garcia-Mila, Zohar, & Andersen, 1995) characterised this process of inquiry learning as co-ordination between theory and evidence. That is, by generating hypotheses, designing and conducting experiments, and inferring conclusions, students construct, expand, and restructure their subject-matter knowledge. In the example of Chen and Klahr (1999), children may have naive theories about the relation between wire diameter and stretching: the thicker the spring the longer the stretch. They may test this hypothesis by comparing two strings with different wire diameter and otherwise equal characteristics. Outcomes may contradict the original hypothesis and the validity of the outcomes has to be evaluated independently from the hypothesis at stake. Kuhn (1988) considered the ability to evaluate evidence by taking various theoretical perspectives without a personal bias as the hallmark of full-fledged co-ordination of theory and evidence. According to Kuhn (1988), three abilities contribute to such a high level of inquiry learning: (1) Thinking *about* theories, not just thinking *with* theories. Recognizing the possibility that a theory may be erroneous. (2) Encoding and representing evidence independently from theoretical perspectives. (3) Evaluating evidence without being biased by a personal point of view.

This study aimed at designing a training programme for children to foster the development of inquiry learning skills. Because Case (1992; see below) proposed that important improvements in inquiry learning skills begin to emerge at the age of 11 and 12 years, we decided to focus our efforts on 11- and 12-year-old children. Before introducing our training programme, problems with the complex task of co-ordinating theory and evidence will be discussed.

First of all, learners have to develop the epistemological understanding that theory and evidence are to be distinguished. According to Kuhn (2000) many children in the age range of 11 and 12 years have absolute views on the world. They are not able to raise various hypotheses to explain ambiguous events or phenomena but stick to one explanation which they consider as absolutely true. In fact, they do not make a distinction between theory and evidence. Although the basic insight that theories originate in the human mind and are not necessarily a truthful representation of reality arises around the age of four, the epistemological insight that theories and evidence are distinct categories has to be discovered time and again in the various domains of knowledge and experience children encounter in the course of their cognitive development. Children may retain a hypothesis despite disconfirming evidence because the hypothesis has been advocated by, for instance, a significant adult. In a similar vein, they may reject an hypothesis although the evidence collected confirms the validity of the hypothesis. In a simulation environment showing the influence of environmental factors on the growth of plants experiments may reveal that insecticides have no influence on the height of a plant. Children may discard this evidence because their parents or teachers have told them that insecticides stimulate the growth of plants.

A second problem learners may encounter in a computer-supported learning environment is their inability to simultaneously deal with two or more independent variables when formulating hypotheses, designing experiments or interpreting experimental evidence. Siegler (1978) analysed childrens' reasoning processes in the context of the balance scale problem of Inhelder and Piaget (1958). A balance scale contains a varying number of weights placed on pegs at various distances from the fulcrum. The arms of the balance are kept in a fixed position and the child is asked to predict which arm will go down as soon as the arms are released. Children of age 4 and 5 either randomly guess or count the number of weights on each arm of the scale, disregarding the distance from the fulcrum at which the weights are placed. Full understanding of the interaction between weight and distance does not arise normally before the age of 16. Case (1992) made a relevant distinction between four levels of cognitive development: (1) predimensional level (children refer to global aspects, and only salient differences between alternatives give rise to appropriate answers); (2) unidimensional (children focus on one dimension); (3) bidimensional (children refer to both dimensions but do not quantify differences); (4) bidimensional, with elaboration (children are able to quantify the difference between two dimensions). Case's (1992) distinction between levels of dealing with multidimensional problems illustrates the second problem in inquiry learning, causing learners in computer-supported simulation environments to overlook important relations between variables.

The problem of simultaneously dealing with two or more independent variables becomes manifest when learners have to *design and conduct experiments* in computer-supported learning environments. Siegler and Liebert (1975) showed that 10-year olds are less able to generate all possible combinations of variables in 2×2×2×2 factorial experiment than 13-year olds, even when they are offered information with regard to the possible number of combinations. At a more detailed level, children inadequately control the manipulation of independent variables, for instance by varying only one thing at a time (VOTAT; Tschirgi, 1980).

Lack of skills and/or strategies for designing and conducting experiments make computer-supported simulation environments for inquiry learning often ineffective. The question is whether children can be trained to become real scientists, equipped with appropriate skills and strategies for benefiting from computer-supported simulation environments. Apart from teaching inquiry learning skills, most training programmes provide opportunities for practice. Several examples of training and support procedures exist in the literature (see De Jong & Van Joolingen, 1998, for an overview). Most of these examples contain on-line facilities like a hypothesis scratchpad or an assistant providing solicited or unsolicited help. In other computer-supported simulation environments students are led through a series of increasingly complex models behind the simulation environment (White & Frederiksen, 1990). Model progression facilities tune computer-supported simulation environments to the student's current level of mastery. Metz (1995) suggested that children are able to develop the necessary skills for inquiry learning by offering them authentic experiences in goal directed tasks. This can be accomplished in simulation environments, by providing scaffolding or by arranging collaborative learning tasks. Practice has been shown to improve learning outcomes in inquiry learning tasks (Kuhn, Garcia-Mila, Zohar, & Andersen, 1995; Glaser, Schauble, Raghavan, & Zeitz, 1992; Kuhn & Angelev, 1976; Siegler, Liebert, & Liebert, 1973).

In this study, training inquiry learning skills and providing opportunities for practice were contrasted by comparing the effects of a training programme to foster inquiry learning skills with a practice condition in which students gained experience in conducting experiments in a computer-supported inquiry learning environment without instructional support. The training programme dealt with two strongly intertwined domain-independent strategies related to successful inquiry learning. In the first place, we focused on fostering the epistemological understanding that theory and evidence are to be distinguished by training students to interpret the results of experiments. Successful learning outcomes depend on correct interpretation of the results of the experiments generated. A learner needs to know what can and cannot be inferred from the results of experiments. Therefore, we trained learners to validly interpret the results of experiments in terms of main effects, irrelevant effects and especially interaction effects. In the second place, we fostered to ability to simultaneously deal with two or more independent variables when formulating hypotheses, designing experiments or interpreting experimental evidence by training our learners in consistently applying the CVS (Control-of-Variables Strategy; Chen & Klahr, 1999). Usage of the CVS is a prerequisite for the valid interpretation of experimental outcomes.

In summary, this study focused on the effects of training and practice on inquiry learning in a computer-supported simulation environment. Two groups were compared: a training group and a practice group. Training and practice effects were tested in two computer-supported simulation environments in the domains of biology and geography. In this way, domain effects could be taken into account. Our hypothesis was that the training group would have a better learning outcome than the practice group, irrespective of type of domain. Because we trained our learners to validly infer main effects, irrelevant effects as well as interaction effects, we expected this improvement to be general across all effects in the tasks. Training and practice effects were tested by administering two inquiry learning tests in the simulation environment and comparing outcome measures (comprehension of the underlying model) and process measures (applying the CVS, arriving at valid inferences by interpreting results from experiments in the simulation environment).

1. Method

1.1. Participants

Sixty-two sixth-grade children (30 boys, 32 girls, mean age: 11 years, 4 months, S.D.: 6 months, range: 10 years, 4 months–12 years, 11 months) participated in this study. The children belonging to the training group ($n=31$) came from a primary school in Leiden. The children from the practice group ($n=31$) came from a primary school in Amsterdam. Both schools had a history of high percentages of children continuing their education in the highest level of the Dutch educational system. Before the children participated in the study, their parents had given their written consent. They were informed about the purpose of the study and were told that they would receive a brief overview of the results when the study was finished. To control for differences in cognitive ability between the two groups, a Number Series test (Elshout, 1976) was administered. A Number Series test was chosen as an indicator of reasoning ability, which, according to Carroll (1993), is the central component of fluid intelligence (Cattell, 1987), and which was assumed to be an important predictor of the participants' performance during the inquiry learning tasks. The Number Series test contained extrapolation problems: "Given the series 3 6 5 10 9 ? Which number should be put on the place of question mark? Select the correct answer from the following alternatives: 8 10 14 18 21." The correlation between the test Number Series and learning outcome (combined comprehension scores for biology and geography) amounted to 0.38 ($P<0.01$, $n=57$). No difference in score on the test Number Series was found between the training group ($M = 17.9$, S.D. = 4.0) and the practice group ($M = 17.4$, S.D. = 5.3, $t\,(57) = -0.373$, $P = 0.71$).

1.2. Materials

1.2.1. Inquiry learning tasks

In the domain of *biology*, two inquiry learning tasks were designed, the Plant Growing Task and the Food Task. In the *Plant Growing Task* (see Fig. 1) the objective was to investigate how the growth of a plant, measured by the height of the plant, was influenced by five environmental factors. The independent variables were: (1) giving water, either once or twice a week; (2) usage of an insecticide to keep away plant louses or not; (3) putting some dead plant leaves in the flower pot or not; (4) placing of the plant, either indoors, on a balcony or in a greenhouse and (5) size of the flower pot, either big or small. The levels of the dependent variable (height of the plant) were: 5, 10, 15, 20 and 25 cms. Variable 1 and 5 interacted and variable 2 and 3 were irrelevant. Variable 4 had a main effect. Two of the three levels produced the same effect on the dependent variable, and one level produced a deviating effect. This main effect will be call "curvilinear" henceforth. In the *Food Task* the problem was to find out how eating and drinking habits affected the health status of an imaginary person, called Hans. The independent variables were: (1) snacks, either fried cakes (Dutch specialty), ice cream or pastry; (2) carbohydrates, either cornbread, potatoes and pasta or white bread, potatoes and rice; (3) alcohol, a glass of wine a day or not; (4) albumen, either steak, filet of chicken and low-fat milk or chops, drumsticks and normal milk and (5) vitamins, either taking a vitamin supplement or not. The health status was a five-point

Fig. 1. Interface of the FILE simulation environment for inquiry learning.

scale ranging from 1 (very unhealthy) to 5 (very healthy). Variable 2 and 4 interacted, variable 1 had a main, curvilinear effect and 3 and 5 were irrelevant.

In the domain of *geography*, again two inquiry learning tasks were designed, the Otter Task and the Aging Task. In the *Otter Task* participants had to find out how different factors affected the number of otters living in the Netherlands (in the Netherlands, the otter is bound to become extinct). The factors were: (1) food, either providing otters with extra fish or not; (2) environmental pollution, either reducing the pollution to a minimum, an intermediate level or doing nothing to reduce it; (3) natural habitat, either letting otters live in one big area or in several separated smaller areas; (4) media, either making the extinction of the otters known to a large public via television programs or not and (5) either closing the natural surroundings of the otter to the public or not. The size of the otter population could be: 100, 250, 400, 450 and 600 otters. Variable 3 and 5 interacted, variable 2 had a main, curvilinear effect and 1 and 4 were irrelevant. Finally, in the *Aging Task* different factors affected the extent to which the population of a certain country would age. These factors were: (1) state of the economy, either poor or good; (2) quality of the educational system, either poor or good; (3) most important means of living, either business, industry or agriculture, (4) climate, either rainy or sunny and (5) general safety, either safe or unsafe. The extent to which the population aged in different situation was measured on a five-point scale ranging from 1 (low level of aging) to 5 (high level of aging). Variable 1 and 2 interacted, variable 3 had a main, curvilinear effect and 4 and 5 were irrelevant.

1.2.2. Inquiry Learning Environment

The computer-based learning environment FILE (Hulshof, Wilhelm, Beishuizen, & Van Rijn, in press; Wilhelm, Beishuizen, & Van Rijn, in press) has been designed for studying the process of scientific discovery. Students can design and run experiments by manipulating five independent variables and observing the effects on one dependent variable. The independent and dependent variables of each of the four Inquiry Learning Tasks have been explained above. The Inquiry Learning Environment will be explained by taking the Plant Growing task as an example. The interface is shown in Fig. 1.

Participants conduct an experiment by choosing values for each of five independent variables. In Fig. 1, two experiments have been conducted. Independent variables are depicted in the left margin of the screen. Each variable has two or three levels. Students choose a level by clicking on the appropriate pictogram. Pictograms show the choice of the independent variables in the large right area. During the first experiment the participant has selected (1) “once a week” for the independent variable “giving water”, (2) “no insecticide” for the variable “usage of an insecticide to keep away plant louses”, (3) “no leaves” for the variable “putting some dead plant leaves in the flower pot”, (4) “indoors” for the variable “placing of the plant”, and (5) “small” for the variable “size of the flower pot”. In Fig. 1, the participant is preparing experiment 3 by choosing values of the independent variables. Three values have already been specified. Before actually running an experiment, participants may be requested to specify an expected outcome. In the first experiment, the expected outcome was 5, in the second experiment it was 15. Conducting an experiment by pressing the “Resultaat” (“Result”) key produces a value of the dependent variable, in this case the height of the plant. In Fig. 1 both experiment 1 and experiment 2 resulted in a height of “20”. Participants continue with carrying out experiments until they feel satisfied and are able to specify the relation of all independent variables with the dependent variable. All keystrokes and mouseclicks are recorded, together with a time stamp.

The model behind the simulation can be defined by the instructor. In the experiment reported here, two independent variables were irrelevant, two variables interacted and one variable (with three levels) had a curvilinear relationship with the dependent variable.

1.2.3. Inquiry learning training

Children in the training group received an individual one-hour training in which they learned how to interpret sets of experiments in terms of main effects, irrelevant effects and interaction effects of independent variables. The training consisted of two parts. In the first part, skills in inferring main effects, irrelevant effects and interaction effects from pre-selected experiments were trained. In the second part, skills in conducting the proper experiments to test these effects were the focus of attention. The second part was included to bridge the gap between inferring effects from pre-selected experiments and testing effects with self-selected experiments.

The first part of the training consisted of 35 training instances. Each instance was presented on paper and contained a set of experiments (two, three or four) and their outcomes. The outcomes were covered during the first and initial part of the second step of the instructional dialogue. The way experiments were represented was identical to the way they were represented in the inquiry learning tasks. However, in the learning tasks there were five independent variables and during the training session the experiments consisted of four independent variables. The children were trained in two everyday-life domains (shopping in a supermarket and driving to school). Each instance was handled according to a fixed instructional dialogue consisting of four steps. At each step, the trainer asked standardized questions and gave feedback about the correctness of the answers. At the first step, the instance was presented with the outcomes covered and the trainer asked the child to name the pictures in each row (''Could you name the pictures in each row?''). At the second step, the trainer asked: ''If you look at these rows, what can you learn from them about what makes a difference?'' The purpose of this question was to see if the children could pinpoint the independent variable(s) of which the effect(s) could be validly tested given the set of experiments. At this point, the experimenter could check whether the children understood the CVS by making inferences about variables which were adequately manipulated. If they did not, the experimenter would explain to them the principle of the CVS. Consequently, the children were allowed to uncover the outcomes. At the third step, the trainer asked: ''What can you learn from these outcomes about what makes a difference?'' The children then could check the effect(s) of the variable(s) chosen during the second step and infer main effects, irrelevant effects, a combination of both, or interaction effects. When the answer was correct the trainer gave feedback and summarized the findings (fourth step). If children answered the trainer's questions correctly, subsequent instances could be skipped. If not, new instances of a particular effect were presented, until the child correctly identified the effects. In this way the length of the training was adapted to the individual child.

During the second part of the training, the trainer offered the children little pictures of the levels of the independent variables and asked the children to show how a main effect could be tested. When the experiments were designed correctly (i.e., the child had to make two identical rows of pictures and manipulate the independent variable), the experimenter would place two cards with outcomes at the end of the rows making sure a main effect could be inferred. In the same way, the children were asked to show how an interaction effect could be tested. Conducting experiments to test an irrelevant effect was omitted because this is in essence identical to testing a main effect.

1.2.4. Inquiry learning tasks

After the training session two inquiry learning tasks were presented to test training effects. Participants in the practice condition received two inquiry learning teasks, both during the first and during the second practice session. A learning task was ended when the learners gave notice that they had completely studied all relations between the independent variables and the dependent variable. After finishing a task, the learners were asked what effect each independent variable had on the dependent variable. This interview was later scored resulting in a *comprehension score* for the test. The effects of the independent variables on the dependent variable can be summarized in a total of nine statements. For example, in the Plant Growing task the correct statement for the effect of using an insecticide to keep away plant louses (irrelevant) was: "Using an insecticide or not does not make a difference." For each correct statement a learner received two points. Zero points were given if the statement was incorrect. A learner could also receive one point when the necessary restrictive conditions in statements about interacting variables were not stated. For example, if the complete description of an interaction was "As long as the plant grows in a big flower pot, giving water once or twice a week does not make a difference", one point was give when the learner stated: "Giving water once or twice a week does not make a difference." The maximum comprehension score for each task was 18 points. Comprehension scores of the two tests administered after the training session and during the practice sessions were added. The combined comprehension score had a minimum of 0 and a maximum of 36. Apart from the comprehension score, the following *learning process measures* were collected: (1) number of variables changed per experiment (indicative of usage of the CVS); (2) mean number of (valid and invalid) inferences per experiment; (3) inference ratio: number of valid inferences divided by the total number of inferences per experiment.

1.3. Design and procedure

The training group consisted of 31 participants. The training session took place in a separate room in the school. The experimenter briefly explained the purpose of the study. The participants were informed that they were going to complete a training and a test session, the second one would follow a couple of days later in the course of the week. During the test session, all participants completed two learning tasks in the same domain, either geography (15 participants) or biology (16 participants).

The practice group consisted of 31 participants. On two separate occasions, the participants completed two practice tasks. For 16 participants the first practice session consisted of two geography tasks and the second practice session contained two biology tasks. For 15 participants, the first practice session contained two biology tasks and the second practice session contained two geography tasks. The practice sessions took place in a separate room in the school. The experimenter briefly explained the purpose of the study. The participants were informed that they were going to complete two sessions, the second one would follow a couple of days later in the course of the week.

We ran two separate ANOVAs to establish training effects. The first ANOVA compared the performance of the training group during their test session with the performance of the practice group during their first practice session. The second ANOVA compared the performance of the training group during their test session with the performance of the practice group during their second practice session. A repeated measures analysis was used to test within subjects differences between the first and the second practice session.

2. Results

2.1. Effects of training and practice on inquiry learning outcomes

To test the differences in learning outcome between the training group and the practice group (first session; see Table 1), a 2 (condition; training vs. practice)×2 (type of domain; biology versus geography) ANOVA was executed. A significant condition effect (training versus practice) emerged [F (1, 57) = 7.20, $P < 0.05$]. The training group ($m = 20.9$, S.D. = 6.2) outperformed the practice group during their first session ($m = 16.9$, S.D. = 4.9). No main effect for type of domain or interaction effect was found. A 2×2 ANOVA with the same factors was used to test the difference in learning outcome between the training group and the practice group during their second session. No significant main effects or interaction effects were detected. Differences in learning outcome between the first and the second session of the practice group were tested with a repeated measures analysis with test session (first versus second session) as a within-factor and task order (biology—rgeography versus geography—biology) as a between-factor. A significant main effect for test session emerged [F (1, 26) = 32.64, $P < 0.001$]. Performance during the second session ($m = 20.8$, S.D. = 3.1) was better than performance during the first session ($m = 16.9$, S.D. = 4.9). No significant order effect or interaction effect was detected. Thus, it appears that across domains, learning outcome of the practice group during their first session was inferior to performance during the second session and to performance of the training group. Their was no difference between the performance of the practice group during their second session and the training group.

The above analyses were also performed to test the effects of training and practice on detecting curvilinear, irrelevant and interaction effects. As far as detecting curvilinear effects is concerned, across domains, the difference in separate comprehension scores for this particular effect between the training group and the practice group during their first session or their second session was

Table 1
Mean comprehension scores of the training group and the practice group[a]

	Training $n = 31$		Practice $n = 31$			
	M	S.D.	First practice session		Second practice session	
			M	S.D.	*M*	S.D.
Geography	20.7	5.9	17.5	5.0	20.5	3.0
Biology	21.1	6.7	16.2	5.0	21.2	3.3
Total	20.9	6.2	16.9	4.9	20.8	3.1

[a] Maximum comprehension score is 36 (summarized over two tasks).

not significant, but the difference between the two sessions of the practice group was significant [F (1, 26) = 11.85, $P < 0.001$]. No significant interaction effects were found. With respect to the separate comprehension scores for the irrelevant effects, the across-domain difference between the training group and the practice group during their first session reached significance [F (1, 57) = 7.85, $P < 0.01$], the difference between the training group and the practice group during their second session was not significant and the difference between the two sessions of the practice group was significant [F (1, 26) = 27.80, $P < 0.001$]. No significant interaction effects were found. With respect to the separate comprehension scores for the interaction effect, all comparisons between and within the groups were non-significant. Means and standard deviations of the interaction comprehension scores (maximum score is 8) for the training group and the first and second session of the practice group were 4.6 (2.6), 3.9 (.9) and 4.3 (1.2), respectively.

This result did not support our initial hypothesis with respect to the expected group differences in discovering the interaction effect. However, informal observations showed that several children in the training group referred to the training items in which interaction effects could be discovered when they manipulated the interacting variables in the learning tasks. Therefore, we performed an additional analysis. A complete description of an interaction effect relates the value of the dependent variable to the interacting independent variables. An example of such a statement for the plant growing task is: "In a large flower pot, giving water once or twice does not make a difference." This statement was awarded with two points. When a learner gave an incomplete statement: "Giving water once or twice does not make a difference", he or she was awarded with one point. We re-analyzed differences between and within groups using the mean number of complete statements, awarded with two points. Across domains, in the training group 26 of such statements were given by 10 children ($m = 0.90$, S.D. = 1.45). During the second session of the practice group, four of these statements were given by one learner ($m = 0.13$, S.D. = 0.72) and during the first session of the practice group four of these statements were given by three learners ($m = 0.13$, S.D. = 0.43). The difference between the training group and the practice group during their first session as well as the difference between the training group and the practice group during their second session reached significance [F (1, 58) = 7.98, $P < 0.01$, F (1, 58) = 6.90, $P < 0.05$, respectively].

For each of the two domains (biology and geography) separately, comparisons were made between the training group and the practice group. Within the training group, 16 participants were tested in the biology domain. Within the practice group, 15 participants started with two practice tasks in the biology domain whereas the remaining 16 participants carried out these two practice tasks in the biology domain during their second session. The learning outcomes of these three groups differed significantly [F (2, 41) = 4.17, $P < 0.05$]. Post hoc tests (Tukey) showed that both the training group ($m = 21.1$, S.D. = 6.7) and the participants who conducted the biology tasks during their second practice session ($m = 21.2$, S.D. = 3.3) had a significantly higher comprehension score than the participants who completed the biology tasks during their first practice session ($m = 16.2$, S.D. = 5.0). In the domain of geography no significant differences in learning outcome between the groups were found.

2.2. *Effects of training and practice on the process of inquiry learning*

We used the mean number of variables changed per experiment (indicative of usage of CVS), the mean number of inferences per experiment, and the ratio between the number of valid inferences and total number of inferences as dependent variables to test our prior hypotheses. In Table 2, the means and standard deviations of the process measures are depicted. In each group, between domain differences in process measures appeared to be absent. T-tests for each group showed no significant differences in process measures between domains. All across domain differences in the learning measures between the training and the practice group during their first session were significant [CVS: $F\ (1,\ 57)=29.88$, $P<0.001$, number of inferences: $F\ (1,\ 57)=50.32$, $P<0.001$, inference ratio: $F\ (1,\ 56)=11{,}67$, $P<0.01$]. For the differences between the training group and the practice group during their second session this pattern was the same, except for the difference in inference ratio, which was not significant [CVS: $F\ (1,\ 58)=4.96$, $P<0.05$, number of inferences: $F\ (1,\ 58)=8.72$, $P<0.05$]. The differences between the two sessions of the practice group showed the same pattern [CVS: $F\ (1,\ 28)=37.99$, $P<0.001$, number of inferences: $F\ (1,\ 28)=35.98$, $P<0.001$ inference ratio: $F\ (1,\ 27)=19.08$, $P<0.001$].

In the domain of geography differences between the training and the practice group showed up with respect to the number of inferences (see Table 2). With respect to the use of CVS, the difference between the training group and the practice group during their first session and between the training group and the practice group during their second session reached significance ($P<0.05$). No differences emerged between the first and the second session of the practice group.

In the domain of biology, differences between the training and the practice group emerged with respect to all the measures (see Table 2). The differences between the training group and the practice group during their first session, and between the training group and the practice group during their second session reached significance ($P<0.05$). No differences emerged between the first and the second session of the practice group.

The correlations between learning process measures measures and learning outcome across domains for the training and the practice group are depicted in Table 3. As can be seen, all measures significantly correlated with learning outcome.

Table 2
Mean comprehension scores of the training group and the practice group

	Training $n=31$		Practice $n=31$			
	M	S.D.	First session		Second session	
			M	S.D.	*M*	S.D.
Use of CVS (max = 5)						
Biology	1.5	0.32	20.3	0.35	10.8	0.37
Geography	1.6	0.56	20.1	0.52	10.8	0.33
Number of inferences (max = 1)						
Biology	0.80	0.14	0.49	0.15	0.72	0.15
Geography	0.83	0.16	0.52	0.20	0.67	0.15
Inference ratio (max = 1)						
Biology	0.87	0.13	0.54	0.21	0.82	0.17
Geography	0.78	0.34	0.66	0.27	0.70	0.17

Table 3
Correlations between learning process measures and across domain comprehension score for the training and the practice group (first session)

	1	2	3	4
1. Comprehension Score	–	0.65*	0.47*	0.55*
2. Usage of CVS		–	−0.52*	−0.77*
3. Number of Inferences			–	0.43*
4. Inference Ratio				–

* $P < 0.05$.

3. Discussion

This study contrasted the effects of training and practice to improve inquiry learning skills. The results showed that a training procedure which focuses on the process of evaluation of evidence and the CVS can improve inquiry learning outcome in sixth-grade children. Across domains, the trained learners were better able to identify irrelevant and main effects of variables than learners from a practice group. The practice group showed a comparable improvement in learning outcome as the training group. However, the learners in the training group showed a better understanding of the interaction effects than the learners in the practice group, which implies a benefit of training over the effect of practice. Inspecting these results separately for each domain showed that the improvement in learning outcome could be attributed especially to practice and training effects in the domain of biology. In the domain of geography no differences in comprehension score were found between the two groups. Thus, both training and practice can improve learning outcome but, compared to the practice group, learners benefit from a training which focuses on the discovery of an interaction effect. Furthermore, the practice and training effects appear to depend on type of domain.

So, training and practice appears to improve learning outcomes, but did training and practice result in changes in learning behavior? The answer is yes. Across domains, there were improvements in the use of CVS, the number of inferences, and the inference ratio. Again, these changes were domain-dependent. In the domain of geography, the number of inferences made improved in the training group, compared to the practice group. No differences between the two sessions of the practice group showed up. In the domain of biology, all measures, were significantly higher in the training group, relative to the practice group during their first session. Again, no differences between the two sessions of the practice group showed up.

What could be the reason that no training and practice effects with respect to learning outcome were found in the domain of geography? The skills we focused on in the training were domain-independent, which means that they should be equally applicable in either domain. From Table 2 it is clear that, between domains, the learners did not differ in the inquiry learning process measures. This means that the content of the tasks can only explain the differences found. One hypothesis is that children had less prior knowledge about the content of the geography tasks than about the content of the biology tasks. It might be that the geography tasks were too abstract for the children. Evidence for this hypothesis can be inferred from the correlations between the learning outcomes in both domains and the inquiry learning measures of generating hypotheses and usage of CVS. Wilhelm and Beishuizen (2003) found that when the task domain was abstract, inquiry learning process measures accounted for more variance in learning outcome. When the tasks in the geography domain have an abstract quality, the correlation between generating hypotheses and usage of CVS should be higher in this domain than the same correlations in the biology domain. This was the case. The correlations between generating hypotheses, usage of CVS and learning outcome in the geography domain were 0.38 ($P<0.05$) and −0.69 ($P<0.001$), respectively. The corresponding correlations in the biology domain were 0.04 ($P=0.812$) and −0.49 ($P<0.01$). These results are compatible with the results of Wilhelm and Beishuizen (2003) who found similar differences in tasks of concrete and abstract content. In tasks of abstract content, generating hypotheses and the use CVS accounted for more variance in learning outcome than in tasks of concrete content. Thus, the reason why learning outcome in the geography tasks remained the same despite training or practice might be attributed to the abstract, or unfamiliar content of the domain. It seems that the benefit of practice and training in terms of learning outcome depends on the concreteness, or familiarity of the task domain in which the effects of training and practice are tested. In familiar, concrete domains students possess prior knowledge on which they are able to successfully apply inquiry learning skills acquired during the training session. The data suggest that acquired inquiry learning skills can only be transferred to familiar domains, in which the learner can activate prior knowledge.

The question whether we can train sixth-grade children to behave like scientists can be answered in an affirmative way. We agree with Metz (1995) that presumed Piagetian developmental constraints are not tenable in the sense that science instruction at the end of the elementary school should be based on them. The data of this study show that both training and practice lead to the desired changes in learning outcome and learning behavior. However, it seems that training and practice work in familiar domains, but not in unfamiliar domains. Although the question remains whether it is familiarity per se which produced the effects found (see Wilhelm & Beishuizen, 2003), the results imply that for appropriate science instruction, the level of familiarity of a task domain should be taken into account. In familiar domains, it is easier for learners to apply inquiry learning skills they already possess or are trained in. In unfamiliar domains, learners can acquire knowledge by consciously applying experimenting skills. However, learners should be introduced to research methods in familiar domains. When the application of experimenting skills has reached an adequate level, less familiar domains can be introduced. This means that teachers need to carefully choose the domains in which inquiry learning tasks used for science instruction are embedded. When the domain is unfamiliar, it might be better to introduce learners to this domain first, before the learners start experimenting, particularly when the new domain contains complex effects like interactions between independent variables which might be too difficult for learners to comprehend.

References

Carroll, J. B. (1993). *Human cognitive abilities. A survey of factor-analytic studies*. Cambridge: Cambridge University Press.

Case, R. (1992). *The mind's staircase. Exploring the conceptual underpinnings of children's thought and knowledge*. Hillsdale, NJ: Erlbaum.

Cattell, R. B. (1987). *Intelligence: its structure, growth, and action*. Amsterdam: North-Holland.

Chen, Z., & Klahr, D. (1999). All other things being equal: acquisition and transfer of the control of variables strategy. *Child Development*, *70*(5), 1098–1120.

De Jong, T., & Van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, *68*(2), 179–201.

De Jong, T., Beishuizen, J. J., Van Someren, M., Veenman, M., Hulshof, C., Wilhelm, P., Prins, F., & Van Rijn, H. (in press). Determinants of discovery learning in a complex simulation environment. In P. Gärdenfors, U. Riss, & P. Johansson (Eds.). *Cognition, Education and Communication Technology*. Hillsdale, NJ: Erlbaum.

Elshout, J. J. (1976). *Karakteristieke moeilijkheden in het denken dissertation*. Amsterdam: University of Amsterdam.

Glaser, R., Schauble, L., Raghavan, K., & Zeitz, C. (1992). Scientific reasoning across different domains. In E. De Corte, M. C. Linn, H. Mandle, & L. Verschaffel (Eds.), *Computer-based learning environments and problem solving* (pp. 345–371). Heidelberg: Springer Verlag.

Hulshof, C. D., Wilhelm, P., Beishuizen, J. J., Van Rijn, H. (in press). FILE: A tool for the study of inquiry learning. *Computers in Human Behavior*.

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

Kuhn, D., & Angelev, J. (1976). An experimental study of the development of formal operational thought. *Child Development*, *47*, 697–706.

Kuhn, D. (1988). Summary and conclusions. In D. Kuhn, E. Amsel, & M. O'Loughlin (Eds.), *The development of scientific skills* (pp. 219–236). New York: Academic Press.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, *96*(4), 674–689.

Kuhn, D. (2000). Theory of mind, metacognition, and reasoning: a life-span perspective. In P. Mitchell, & K. J. Riggs (Eds.), *Children's reasoning and the mind* (pp. 301–326). Hove, UK: Psychology Press.

Kuhn, D., Garcia-Mila, M., Zohar, A., & Andersen, C. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development*, *60*(4), 1–128 Serial No.

Metz, K. (1995). Reassessment of developmental constraints on children's science instruction. *Review of Educational Research*, *65*, 2 93-127.

Siegler, R. S. (1978). The origins of scientific reasoning. In R. S. Siegler (Ed.), *Children's thinking: What develops?* (pp. 109–149). Hillsdale: Erlbaum.

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

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development*, *51*, 1–10.

White, B. Y., & Frederiksen, J. R. (1990). Causal model progressions as a foundation for intelligent learning environments. *Artificial Intelligence*, *42*, 1 99-157.

Wilhelm, P. (2001). *Knowledge, skills and strategies in self-directed inductive learning*. Unpublished doctoral dissertation. Leiden: Leiden University.

Wilhelm, P., & Beishuizen, J. J. (2003). Content effects in self-directed inductive learning. *Learning & Instruction*, *13*, 381–402.

Wilhelm, P., Beishuizen, J. J., & Van Rijn, H. (in press). Studying inquiry learning with FILE. *Computers in Human Behavior*.