ELSEVIER

Learning and Instruction 18 (2008) 580–592

Learning and Instruction

www.elsevier.com/locate/learninstruc

The influence of domain knowledge on strategy use during simulation-based inquiry learning

Ard W. Lazonder*, Pascal Wilhelm [1], Mieke G. Hagemans

Department of Instructional Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

Received 8 August 2007; revised 9 October 2007; accepted 6 December 2007

Abstract

This study investigated how students' knowledge of a particular domain influences the type of investigative strategy they utilize in an inquiry learning task within that domain. Students with high domain knowledge were assumed to employ a theory-driven strategy, whereas less knowledgeable students were expected to start off in a data-driven mode of inquiry and gradually shift to a theory-driven strategy. Participants were 36 college freshmen who performed a simulation-based inquiry task in a familiar domain and another, isomorphic task, in an unfamiliar domain. Within-subject comparisons of the hypotheses participants generated on both tasks generally confirmed predicted patterns of strategy use. Results further showed that participants performed the familiar task more successfully. Implications for inquiry learning environments are discussed.

Keywords: Inquiry learning; Discovery learning; Computer simulations; Prior knowledge

1. Introduction

''I hear and I forget. I see and I remember. I do and I understand''. This ancient Confucius quotation reflects contemporary notions of learning which emphasize that learners should be active agents of their own learning processes (Bransford, Brown, & Cocking, 2002). These educational beliefs coincide with the tenets of inquiry learning, a pedagogy in which learners infer knowledge about a domain by generating hypotheses and designing and executing experiments to validate these hypotheses. Although inquiry learning can lead to a deeper and more meaningful understanding, the effectiveness of this mode of learning is challenged by intrinsic problems many learners have with the complex, integrated set of skills inquiry learning entails (De Jong & Van Joolingen, 1998). Without proper support, inquiry learning is indeed a rather ineffective and inefficient way of learning (Klahr & Nigam, 2004; Mayer, 2004), certainly for learners with little prior knowledge (Tuovinen & Sweller, 1999). Many studies have therefore examined how cognitive support tools can compensate for learners' inquiry skill deficiencies (for an overview see De Jong, 2006a, 2006b). The question of how inquiry learning can be tailored to the learning needs of students with little prior knowledge has received significantly less attention.

* Corresponding author. Tel.: +31 53 489 3082; fax: +31 53 489 2849.
E-mail address: a.w.lazonder@utwente.nl (A.W. Lazonder).
[1] Pascal Wilhelm is now at Saxion Universities of Applied Sciences, The Netherlands.

0959-4752/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.learninstruc.2007.12.001

Research in the area of problem solving has shown that inaccurate or incomplete domain knowledge may impede the learning process. In their review of the literature, Alexander and Judy (1988) concluded that more knowledgeable learners select more sophisticated strategies and execute these strategies more effectively than learners with lower levels of prior knowledge. Other studies have shown that this conclusion generalizes to inquiry learning. For example, Schauble, Glaser, Raghavan, and Reiner (1991) found that learners with more sophisticated initial conceptions of the domain also employed more sophisticated strategies to induce knowledge from the simulation. Lavoie and Good (1988) demonstrated that learners with high initial knowledge were more successful in predicting experimental results. Hmelo, Nagarajan, and Roger (2000) found that learners with high prior domain knowledge displayed well-structured, goal-oriented inquiry behavior, whereas learners with little domain knowledge performed unsystematically. Developmental research has shown that these findings generalize across age levels (e.g., Veenman, Wilhelm, & Beishuizen, 2004). A review by Zimmerman (2007) further showed that the development of inquiry skills follows the same general course in children and adults, and is reciprocally influenced by increases in domain knowledge.

While these studies established how prior knowledge accounts for differences in strategy use, they leave unaddressed the fact that students accumulate knowledge during their inquiries. In order to examine how students' existing *and* evolving knowledge structures affect their choice of investigative strategies, a cognitive model of knowledge acquisition is required which is sensitive to changes in the students' knowledge. Klahr and Dunbar's (1988) model of Scientific Discovery as Dual Search (SDDS) meets this requirement. The model characterizes a student's inquiry learning activities as a search in two related problem spaces: the hypothesis space and the experiment space. The hypothesis space contains a student's knowledge of the relations between the variables in the domain; the experiment space contains all possible experiments that can be conducted with the equipment at hand. Scientific discovery then proceeds in iterative cycles by searching the hypothesis space for a testable hypothesis, searching the experiment space for a combination of experiments to test the hypothesis, and evaluating evidence to verify or refine the hypothesis. Alternatively, if students are unable to retrieve a testable hypothesis from the hypothesis space, they can search the experiment space for exploratory experiments that will help them formulate new hypotheses.

Van Joolingen and De Jong (1997; see also Gijlers & De Jong, 2005) adapted the SDDS model for simulation-based inquiry learning in complex domains. Of particular relevance here is the refinement of the hypothesis space to elucidate how students develop knowledge from their interactions with the simulation. Toward this end Van Joolingen and De Jong (1997) introduced the concept of a universal hypothesis space (containing all possible hypotheses in a domain) and divided it into three subspaces (see Fig. 1). The learner hypothesis space contains all possible hypotheses a student can generate about a domain. The learner domain space embodies the students' beliefs and ideas about the relations between the variables in a domain and, thus, represents the students' knowledge base. The target conceptual model contains the relations between the variables in the simulation students have to induce.

With the extended SDDS model, knowledge building during inquiry learning can be depicted as changes in the learner domain space. The learner domain space initially houses a student's prior knowledge about the task. In the example shown in Fig. 1, there is no overlap between the learner domain space and the target conceptual model, meaning that the student has no prior knowledge about the constellation of variables and relations in the simulation. Since

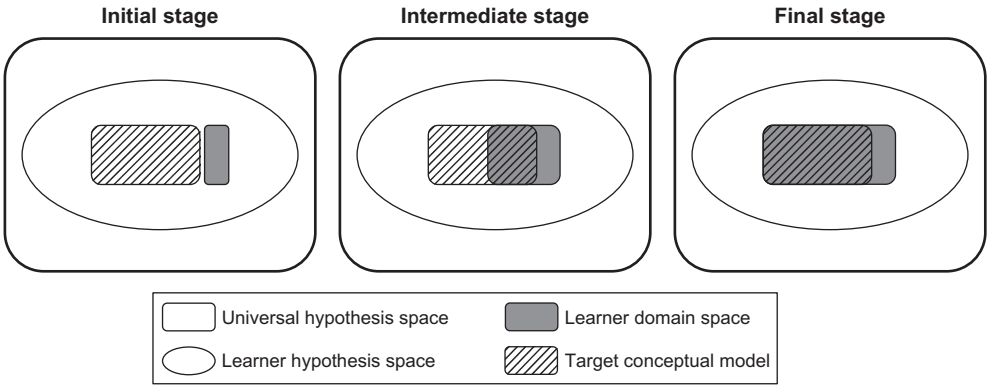

Fig. 1. Example of changes in the hypothesis space during the inquiry learning process. The experiment space, which contains all possible experiments learners can perform with a simulation, remains unchanged through time and is, therefore, not included in this figure.

the target conceptual model lies within the learner hypothesis space, this student is able to generate (and test) relevant hypotheses to infer information about the target conceptual model. Doing so may cause the learner domain space to shift toward the target conceptual model. This may involve adding new propositions to the learner domain space, or removing existing propositions that could not be confirmed through experimentation (i.e., misconceptions). Together these changes denote a student's knowledge gains, which should ideally lead to a complete overlap between the learner domain space and the target conceptual model.

Klahr and Dunbar (1988) used the SDDS model to examine the dynamic interplay between students' domain knowledge and their approach to the inquiry task. Participants in this study were characterized as Theorists or Experimenters based on their experimentation behaviors. Both groups initially used hypotheses to guide the selection of experiments, but their approaches diverged once the initial hypotheses were abandoned. Theorists kept searching the hypothesis space for new hypotheses; Experimenters applied a data-driven approach and explored the experiment space to see if they could induce regularities from experimental outcomes. This seems to suggest that Theorists had more prior knowledge which might have influenced the strength of their initial hypotheses and facilitated their search for alternative hypotheses. This supposition could, however, not be confirmed from the data obtained: although Klahr and Dunbar (1988) identified three possible sources of prior knowledge, they did not assess the participants' initial conceptions of the to-be-discovered phenomenon.

Thus, there appears to be a dichotomization of domain knowledge in the research on investigative strategy use. The work of Lavoie and Good (1988), Schauble et al. (1991), and Hmelo et al. (2000) demonstrated how prior knowledge facilitates effective strategy use, whereas the Klahr and Dunbar (1988) study established how students' evolving knowledge shapes their strategies of investigation and analysis. The present study sought to integrate both lines of research. The study took Klahr and Dunbar's (1988) research as a starting point, and modified their experimental set-up to allow for a more controlled comparison. The main differences concerned the assessment of prior knowledge, and the within-subject comparison of strategy use across two inquiry tasks. For the sake of experimental rigor, the study polarized differences in prior knowledge by having participants perform a concrete inquiry task in a familiar domain and an abstract task in an unfamiliar domain.

The concrete task involved a realistic problem and was designed so that participants should have considerable prior knowledge of three of the four variables. A pretest ascertained that participants who entered the study met this requirement. The fourth variable concerned a ''mystery factor'' that was inspired by the characteristics of the Klahr and Dunbar (1988) task and served to establish a realistic analog to inquiry learning in schools. The abstract task asked participants to discover how four geometrical shapes influenced a numerical score. As both the geometric shapes and their effect on the number of points were arbitrary, participants were assumed to have no prior knowledge of the variables and the task domain.

1.1. Hypotheses

Van Joolingen and De Jong's (1997) extended SDDS model was used to predict within-subject differences in strategy use. On the concrete task, the learner domain space comprised the learner's prior knowledge about the three familiar variables and their initial beliefs about the mystery factor. As the pretest assured that the prior knowledge was ''correct'', the learner domain space largely overlaps the target conceptual model. Learners thus have to infer relatively few new properties; their well-equipped learner domain space further enables them to easily generate and test specific hypotheses about these properties. The students' approach to the concrete task was thus assumed to resemble the approach taken by the Theorists in Klahr and Dunbar's study. In contrast, the learner domain space for the abstract task was empty. Learners were, therefore, expected to start off in an Experimenters' mode by performing exploratory experiments to infer knowledge about the domain. As the newly acquired knowledge becomes part of their learner domain space, learners were assumed to gradually shift from an Experimenters' approach to a Theorists' approach during the intermediate and final stages of their inquiry (hypothesis 1).

These strategy differences were further assumed to yield differences in task performance. Consistent with Klahr and Dunbar (1988), learners were expected to perform the concrete task more successfully and efficiently than the abstract task. As the concrete task involved comparatively less new information, and offered more opportunities for learners to exploit prior knowledge to generate hypotheses, they could select a few key experiments (and would thus require less time) to discover the relations between the variables in the task (hypothesis 2).

2. Method

2.1. Participants

Participants were 40 first-year students in social sciences who volunteered to participate in the experiment for course credits. There were 18 males and 22 females with a mean age of 19.33 years (SD = 1.79). Thirty-two participants were Dutch; eight had the German nationality. All German participants had sufficient command of the Dutch language to be able to understand the verbal instructions and written materials. Four participants did not meet the study's entrance criteria and were removed from the sample; more details appear in the results section. Thus, the final number of participants was 36.

2.2. Inquiry tasks

Participants worked on two simulation-based inquiry tasks that were created in an authoring environment named FILE (Hulshof, Wilhelm, Beishuizen, & Van Rijn, 2005). The concrete task invited participants to investigate how each of four factors influenced an athlete's 10,000 m time. These factors were training frequency (zero, one, or three times per week), smoking (continue or quit), nutrition (sport food, regular food, or junk food), and a ''mystery factor'' called Xelam (yes or no). The baseline score was set at 45 min (zero training sessions, continue smoking, regular food, no Xelam). Outcomes could range from 33 to 51 min depending on the factors' values. Increasing training frequency reduced the 10-km time by 1 min (one weekly session) or 3 min (three weekly sessions). Eating junk food caused a 2-min increase, whereas eating sport food yielded a 2-min decrease in time. Quitting smoking made the athlete to run 4 min faster. The effect of Xelam depended on training frequency. It worsened performance by 4 min if training frequency was zero, had no effect in case of one weekly training session, and improved performance by 3 min if there were three weekly training sessions.

Training frequency, smoking, and nutrition were familiar factors. Most adults know what these factors mean and have some sense of their general direction of effects on the 10-km time. A pretest was administered to check whether this assumption held true for the current sample. The magnitude of these effects on the performance of the athlete in this task was unknown and had to be inferred by experimenting with the simulation. Xelam was the unfamiliar factor. It had an imaginary name and neither its meaning nor its effect on the 10-km time was revealed to the participants: they were told it could be anything from a training program, a detergent, liquor, to a form of meditation.

The abstract task asked participants to discover the influence of four geometrical shapes on a numerical score. Shapes included a triangle (blue or brown), square (orange, purple, or green), circle (red, yellow, or pink), and question mark (black or white). The underlying task structure was copied from the concrete task by replacing its factors by a geometrical shape. Training frequency was replaced by the circle, smoking by the triangle, nutrition by the square, and Xelam by the question mark. The values of the variables were adapted accordingly (e.g., continue smoking = blue triangle; junk food = green square). To preclude possible carry-over effects, the magnitude of effects was changed; outcomes could range from 6 to 18 points. The order of the variables in the simulation interface was changed as well (see Fig. 2).

The operation of the simulation was identical for both tasks. Participants could discover the impact of a single factor by manipulating its values and observing the effect on the output variable (either running time or number of points). Factor values could be set by clicking the corresponding icon on the left side of the screen (see Fig. 2). Selected values appeared in the experiment window on the right. Once all factors were set, participants had to predict the outcome by selecting a value from the pull-down menu. They could then click the Result button to run the experiment. In the experiment window, the actual outcome appeared in boldface; the participants' prediction appeared in roman. Participants could scroll the experiment window to review previously conducted experiments; clicking the magnifying glass button allowed them to inspect a self-selected set of experiments in a separate window. All actions were recorded in a logfile.

2.3. Instruments

A background questionnaire determined the participants' demographic characteristics. A color vision deficiency test was administered to ensure that participants were able to differentiate the differently colored shapes in the abstract

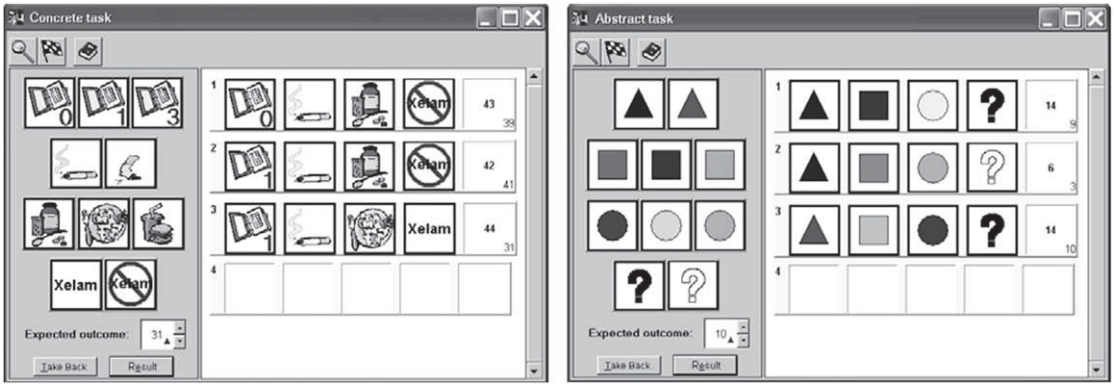

Fig. 2. Simulation interface of the concrete task (left panel) and abstract task (right panel).

task. This test utilized seven items from the Ishihara color test (Ishihara, 1982) that screened participants for red–green deficiencies. Each item contained a circle of colored dots with a digit embedded in a slightly different color that can be read by a person with normal color vision but not by someone with defective color vision. Items were displayed on a 32-bit computer monitor which, although presumably less accurate than the original printed plates, was deemed appropriate for screening purposes. The Kuder-Richardson reliability estimate (K-R 20) of the seven items was 0.95.

A pretest assessed participants' prior knowledge of the concrete task. The 10 items addressing the common factors (i.e., training frequency, smoking, and nutrition) were ''what-if'' questions (Swaak & De Jong, 1996). Each item comprised an initial situation (e.g., ''During an eight-week preparation for a 10-km race, you have trained once a week''), a change (''If you had trained three times per week …''), and three post-change situations that conveyed the possible direction of effects (''… you would probably complete the race [1] faster, [2] just as fast, [3] slower''). Participants answered each item by selecting the most likely post-change situation. Two additional items assessed participants' initial ideas about Xelam. These were open questions that asked participants to write down what they thought Xelam was and how it might affect an athlete's 10,000-m performance.

Content validity of the what-if items was achieved by ensuring representative coverage of the content of the concrete task as well as accurate representation of the variables and relations within that task. Three items addressed the effect of a common factor in general (e.g., ''If you train more often, you will probably complete the race [1] faster, [2] just as fast, [3] slower''). The remaining seven what-if items involved a comparison between two values of the same factor. Every possible combination of values was addressed once; a sample item was given in the previous paragraph. All items dealt with the direction of effects – the magnitude of effects was to be discovered during the experiment – and used the exact same context, factors, values and relations as the concrete task. The internal consistency of the what-if items was satisfactory, K-R 20 = 0.61.

2.4. Procedure

Students participated in the experiment one at a time, receiving the same instructions and following the same experimental procedures. At the beginning of a session, participants completed the pretest and the color vision deficiency test; the background questionnaire was administered when they signed up for the experiment. Next, the experimenter explained the experimental procedures and demonstrated the operation of the simulation by means of a simple inquiry task (determining the costs of a skiing holiday). Participants were then given their first task, using a counterbalanced administration to preclude order effects. Half of the participants were randomly selected to receive the concrete task first and the abstract task second; the other half performed these tasks in reverse order. The concrete and abstract tasks were administered similarly using the procedure outlined below.

Participants started their first task by reading the cover story that introduced the factors in the simulation and asked participants to examine their influence on running time (concrete task) or number of points (abstract task). They then started experimenting with the simulation to investigate the relationships between input variables and outcome

variable. Participants could consult an index card to see how each input variable was visually represented in the simulation interface. They also received an answer form to take notes during the activity and write down their final solution. Participants were allowed 40 min maximum to complete this task. If participants completed the task ahead of time, they would receive their second task. If participants exceeded the time limit, the experimenter interrupted task performance and handed out the second task.

During task performance the experimenter asked participants about their hypotheses. A hypothesis was defined as a statement of the factor under investigation and the presence, direction, or magnitude of its effect on the output variable. While questioning has been criticized for prompting participants to an underlying goal structure for systematic experimentation (Klahr & Carver, 1995), research has shown that non-directive probes have no influence on participants' inquiry learning processes (Wilhelm & Beishuizen, 2004). Hence, two non-directive questions were used to elicit the factor under investigation (''What are you going to investigate?''), and its alleged effect on the output variable (''What do you think will be the outcome?''). These questions were asked every time participants could be testing a hypothesis. That is, questioning occurred whenever a participant (1) had set-up a new experiment and clicked on ''expected outcome'', (2) scrolled the experiment window to review previously conducted experiments, or (3) clicked the magnifying glass button to compare self-selected sets of previously conducted experiments. The experimenter wrote down the participant's responses on a scoring sheet. The reliability of this registration method was assessed by having two raters simultaneously record the responses in five experimental sessions. Analysis of their scoring sheets demonstrated 90% interrater agreement.

2.5. Coding and scoring

Participants' responses to the items of the color vision deficiency test were scored as true or false. Participants passed this test if they could identify the digit in six of the seven items.

Five measures were scored to assess performance and strategy use on the concrete and abstract tasks. These concerned time on task, prior knowledge, and the components of the extended SDDS model (i.e., the experiment space, and the constituent parts of the hypothesis space depicted in Fig. 1). *Time* was assessed from the logfiles. The assessment of the other measures is detailed below. Similar coding and scoring procedures were used for the concrete and abstract tasks unless otherwise stated.

Prior knowledge of the concrete task was assessed by the what-if items of the pretest. These items were multiple-choice questions with three answer options. Participants' answers were checked against the model underlying the concrete task and one point was allocated to each correct response. Two open questions assessed participants' initial conceptions of Xelam. Their responses were coded according to the presumed direction of effect (i.e., positive effect, negative effect, no effect). Although these responses are nothing more but a guess, conjectures too can affect inquiry behavior and should in case of wide divergence be controlled for in the analyses (Wilhelm & Beishuizen, 2003). To ensure that the what-if items were understood correctly, a pilot test was performed with five individuals from the sample population. All five participants achieved the maximum score of 10 points.

The *experiment space* comprised the 36 possible experiments that could be performed with the simulation. Participants' search of the experiment space was scored from the logfiles to indicate the total number of experiments performed. A distinction was made between unique and duplicated experiments. To measure the coverage percentage of the experiment space, the ratio of unique experiments to the total number of possible experiments was computed. The percentage of duplicated experiments was computed by dividing the number of repeated experiments by the total number of performed experiments.

The contents of the *learner hypothesis space* were assessed from the participants' responses to the probing questions. A hierarchical rubric was used to classify the responses (i.e., hypotheses) according to their level of domain specificity. A distinction was made between fully specified, partially specified, and unspecified hypotheses. A fully specified hypothesis comprised two or more factor values and a prediction of the direction *and* magnitude of the effect (''I think the red circle yields a two-point higher score than the yellow circle''). Partially specified hypotheses predicted the direction of effect of two or more factor values (''I think the red circle yields a higher score than the yellow circle''). Unspecified hypotheses merely denoted the existence of an effect (''I think the circle affects the score''). Statements of experimentation plans (''I am going to investigate the circle'') or ignorance (''I have no idea'') were not considered hypotheses. Two raters coded the hypotheses of eight randomly selected scoring sheets of each task. Interrater agreement for both tasks was 0.97 (Cohen's κ).

The contents of the *learner domain space* were assessed from the answer forms. A hierarchical rubric was used to transform these data into a performance success score. Up to 3 points could be earned for each factor, leading to a maximum score of 12 points. Three points were awarded for a factor if both the magnitude and direction of the effect were correct for each value of that factor. Two points were given if the direction of the effect was correct but its magnitude was (partly) incorrect or incomplete. One point was awarded if the answer expressed that a factor affected the output variable, but neither the magnitude nor the direction of this effect was specified correctly. In all cases, ''correct'' was judged from the simulation's underlying model. Two raters used this rubric to score a randomly selected set of eight answer forms of each task. Interrater reliability estimates (Cohen's κ) for the concrete and abstract tasks reached 0.86 and 1.00, respectively.

2.6. Design and data analyses

The first set of analyses concerned a within-subject comparison of participants' performance on the concrete and abstract tasks. One-way repeated measure ANOVA was used to determine the effects of task type on time, performance success, the participants' hypotheses about the variables in the simulations, and the experiments they conducted to generate and test these hypotheses. (Scores on the color vision deficiency test and the pretest served to establish the definitive sample.)

The second set of analyses aimed to provide a more detailed account of the influence of domain knowledge on strategy use. These analyses addressed participants' hypotheses on the concrete and abstract tasks. For each task, a participant's hypotheses were sequenced chronologically. Hypotheses were then divided into quartiles according to their order of appearance, and the average domain specificity of the hypotheses within each quartile was computed. These scores characterize domain specificity as the mean number of constituent elements the hypotheses in a given quartile contained (i.e., presence, direction, and magnitude of an effect; see Section 2.5). Next, sequenced quartile means were analyzed for each participant to identify patterns of how the domain specificity of the hypotheses evolved through time. These patterns represent the approaches participants followed during task performance; *k*-means cluster analysis was performed to determine the average domain specificity of the hypotheses in each approach. Finally, a within-subject comparison was performed to investigate our expectations concerning participants' approaches to the concrete and abstract tasks. Given the relatively small sample size, Fisher's Exact Test was used to assess the interrelationship between participants' approaches on both tasks.

3. Results

The color vision deficiency test and the pretest were used to determine the definitive sample for the analyses. Two male participants did not pass the color test and were removed from the experiment. The remaining 38 participants produced a mean pretest score of 9.13 (SD = 1.27). Their initial conception of the unknown factor was quite consistent: 35 participants thought Xelam would improve the athlete's performance; three participants thought it would have no effect. However, the pretest scores of two participants (one female and one male) were more than two standard deviations below the sample mean. With scores of 4 and 6, respectively, these participants appeared to lack prior knowledge of at least one of the common factors in the concrete task. Both participants were therefore excluded from the sample, which lead to an effective sample size of 36 in the analyses.

On average, these 36 participants needed 27.55 min to complete the concrete task (SD = 10.48). Their mean time on the abstract task was 28.42 min (SD = 11.25). One-way repeated measures ANOVA showed that this difference was not statistically significant, $F(1, 35) = 0.26$, $p = 0.62$.

Participants' performance on both tasks is summarized in Table 1. The experiment space represented the 36 unique experiments that could be performed with the simulations. Participants' search of the experiment space was analyzed in one-way repeated measures ANOVAs. Results showed no significant difference for coverage, $F(1, 35) = 1.47$, $p = .23$, indicating that task type had no effect on the percentage of unique experiments participants performed during their inquiries. There was, however, a significant within-subject difference in the percentage of duplicated experiments, $F(1, 35) = 4.70$, $p < 0.05$, partial $\eta^2 = 0.12$. As the mean scores show, participants repeated experiments less often on the concrete task than on the abstract task.

The content of the learner hypothesis space was analyzed to reveal the nature of the hypotheses that guided participants' search in the experiment space. On the abstract task, over 55% of the participants' utterances were

Table 1
Summary of performance on the concrete and abstract tasks

	Concrete task		Abstract task	
	M	SD	*M*	SD
Experiment space				
Coverage (%)	46.22	15.45	49.15	18.01
Duplicated experiments (%)	14.90	14.38	20.17	18.44
Learner hypothesis space				
No hypotheses (%)	16.92	17.04	56.95	23.68
Unspecified hypotheses (%)	4.18	5.76	4.83	7.90
Partially specified hypotheses (%)	62.61	20.24	12.98	20.49
Fully specified hypotheses (%)	16.51	18.09	25.24	20.48
Learner domain space				
Performance success[a]	10.08	1.99	9.17	2.40

[a] Maximum score = 12; $N = 36$.

not considered a hypothesis according to the definition in the coding and scoring section; on the concrete task this measure approached 17%. One-way repeated measures ANOVA showed this difference to be statistically significant, $F(1, 35) = 105.58$, $p < 0.01$, partial $\eta^2 = 0.75$. Less than 5% of the utterances were classified as unspecified hypothesis. Task type did not yield a significant within-subject effect on this measure, $F(1, 35) = 0.15$, $p = 0.70$. The percentage of partially specified hypotheses was significantly higher on the concrete task than on the abstract task, $F(1, 35) = 137.93$, $p < .01$, partial $\eta^2 = 0.80$; the abstract task showed a significantly higher percentage of fully specified hypotheses than the concrete task, $F(1, 35) = 6.02$, $p < 0.05$, partial $\eta^2 = 0.15$.

The learner domain space contained participants' knowledge of the relations between the variables in the concrete and abstract tasks. A performance success score reflected the extent to which this knowledge matched the simulations' underlying model. Task type had a significant within-subject effect on performance success, $F(1, 35) = 7.52$, $p < 0.05$, partial $\eta^2 = 0.18$, indicating that participants generally were more knowledgeable of the concrete task than the abstract task.

The second set of analyses addressed the changes in domain specificity of the participants' hypotheses on the concrete and abstract tasks. Descriptive statistics are presented in Table 2. Scores in the first quartile indicated that participants' initial hypotheses on the concrete task were more specific than those on the abstract task. Frequency counts revealed that 32 of the 36 participants started the abstract task by performing exploratory experiments that were not guided by a hypothesis. The first hypothesis was generally stated after four of these data-driven experiments were performed ($M = 4.08$, SD = 2.67). On the concrete task, 10 participants started experimenting without a hypothesis. The other participants either used their prior knowledge to generate and test hypotheses about the circumstances under which the athlete would perform best or worst ($n = 23$), or formulate and test a specific hypothesis about one of the variables in the simulation ($n = 3$).

Scores for subsequent quartiles showed that the specificity of the hypotheses on the concrete task remained relatively constant throughout time, whereas the hypotheses on the abstract task gradually became more domain specific. Although these findings are consistent with expectations, the standard deviations on both tasks increased in time, suggesting that the overall pattern in scores may not be observed in all participants. Case-by-case analysis bore

Table 2
Mean domain specificity of participants' hypotheses on the concrete and abstract tasks

	Concrete task		Abstract task	
	M	SD	*M*	SD
First quartile (Q1)	1.74	.52	.47	.53
Second quartile (Q2)	1.97	.61	1.33	.96
Third quartile (Q3)	1.80	.69	1.33	.96
Fourth quartile (Q4)	1.78	.74	1.47	.99

Scores could range from 0 to 3.

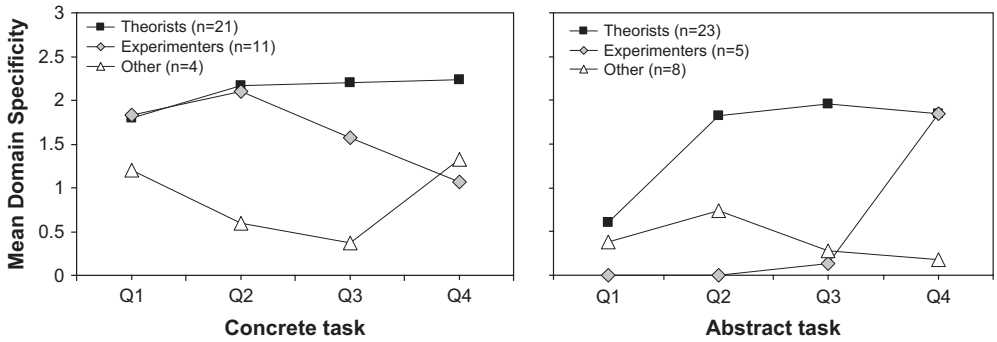

Fig. 3. Classification of participants' approaches to the concrete and abstract tasks on the basis of the domain specificity of their hypotheses in each quartile interval.

this out: from a visual inspection of individual participants' quartile means, three distinct patterns could be identified on the concrete task. *k*-Means cluster analysis of the quartile means was performed to classify participants according to these patterns and determine the average domain specificity of the hypotheses in each pattern through time. The pooled within-cluster SD was 0.52; the Euclidean distance between cluster centers (d) ranged from 1.32 to 2.64. As the left panel of Fig. 3 shows, the hypotheses of 21 participants had high and increasing levels of domain specificity. This is typical of a Theorists' approach in which the contents of the learner domain space are used to generate and test specific hypotheses. Eleven participants followed an Experimenters' approach. As in the Klahr and Dunbar (1988) study, these participants started by formulating and testing specific hypotheses, but gradually switched to a more data-driven mode of experimentation. This can be inferred from the decline in specificity scores, dropping from 2.11 in the second quartile to 1.58 and 1.07 in the third and fourth quartile, respectively. The hypotheses of four participants did not follow a predicted pattern, decreasing from 1.20 in the first quartile to 0.37 in the third quartile and increasing back to 1.33 in the fourth quartile.

Likewise, three patterns were apparent on the abstract task (pooled SD = 0.51; $1.87 \leq d \leq 2.65$). The right panel in Fig. 3 visualizes the progression in scores for each pattern. Consistent with this study's first hypothesis, 23 participants shifted from an Experimenters' approach to a Theorists' approach. They initially performed a few data-driven experiments without statement of a hypothesis, and then started formulating gradually more specific hypotheses, as evidenced by the increase in mean scores for subsequent quartiles. This suggests that the knowledge induced from the initial experiments had become part of participants' learner domain space and was used to generate and test specific hypotheses. Five participants persisted in an Experimenters' approach, conducting data-driven experiments almost until the end of their session. They then generated and tested one specific hypothesis for each variable in the simulation to check whether the knowledge inferred from their previous experiments held true. Eight participants had a mean domain specificity score lower than 1.0 throughout their entire session. These participants apparently failed to infer knowledge from their data-driven explorations, or did not utilize inferred knowledge to formulate testable hypotheses.

Table 3 shows a cross-tabulation of participants' approaches to the concrete and abstract tasks; cell entries represent the number of participants who followed a particular combination of approaches. A significant interrelationship was found between participants' approaches to both tasks, $p < 0.05$, two-tailed Fisher's Exact Test; task order did not

Table 3
Participants' approaches to the concrete and abstract tasks

Concrete task	Abstract task			Total
	Theorist	Experimenter	Other	
Theorist	17	2	2	21
Experimenter	4	3	4	11
Other	2	0	2	4
Total	23	5	8	36

affect this pattern in scores, $p = 0.84$. Frequency counts indicated that 20 of the 34 participants (58.82%), who used a Theorists' or an Experimenters' approach on either the concrete or abstract task, used this approach consistently on both tasks. This was most apparent for the Theorists' approach: 17 of the 21 participants (80.95%), who adopted this approach on the concrete task, also used it on the abstract task. Participants using the Experimenters' approach switched approaches more often: three of the 11 participants (27.27%) who were classified as Experimenter on the concrete task were also considered Experimenters on the abstract task.

4. Discussion

This study investigated the influence of domain knowledge on strategy use during simulation-based inquiry learning. In order to polarize differences in prior domain knowledge, participants performed a concrete inquiry task in a familiar domain, and an isomorphic abstract task they possessed no prior knowledge of. The study's first hypothesis stated that participants would follow a theory-driven approach to the concrete task, and shift from a data-driven to a theory-driven approach in performing the abstract task. As these initial data-driven explorations can be time-consuming and ineffective, the second hypothesis predicted that participants would perform the concrete task more successfully and efficiently than the abstract task. Both hypotheses were generally supported by the results.

Consistent with the second hypothesis, participants attained significantly higher performance success scores on the concrete task (10.08) than on the abstract task (9.17). The height of these scores further indicates that participants inferred new knowledge from the simulation. Knowledge gains on the abstract task are immediately apparent as participants initially were incognizant of the relations between variables in the simulation. The average score on the concrete task exceeded the maximum score participants could have achieved on the basis of their prior knowledge by four points (i.e., knowing the direction of effect of all familiar variables would yield a score of 6 points). Results for measures indicative of the efficiency of participants' performance were mixed. While participants performed relatively fewer duplicated experiments on the concrete task, they needed as much time and performed as many unique experiments on both tasks.

The lack of substantial evidence for the efficiency hypothesis might indicate that participants followed the same approach to the concrete and abstract tasks. This was borne out by the results: strategy use of more than 60% of the participants was consistent across tasks. Theorists in particular maintained their theory-driven mode of inquiry: over 80% applied this approach to both tasks. A closer inspection of their investigative behaviors elucidates why performance on both tasks was equally efficient. Theorists on the abstract task stated their first hypothesis already after four data-driven experiments. By switching to a Theorists' approach during the initial stages of the inquiry, the alleged efficiency differences were minimized.

The hypotheses participants generated on the concrete and abstract tasks are generally consistent with the study's first hypothesis. Surprisingly, however, the highest percentage of fully specified hypotheses was found on the abstract task. This counter-intuitive result may be due to the presence or absence of explanatory mechanisms (Wilhelm & Beishuizen, 2003). On the abstract task, knowledge is built from scratch, and participants generate and test fully specified hypotheses to assess whether newly induced knowledge is correct. This is less pertinent on the concrete task, where prior knowledge serves as framework of reference in assessing experimental outcomes and knowledge. Yet, this method holds a potential danger, as incorrect prior knowledge may cause students to draw incorrect conclusions about experimental outcomes. Although subsequent experiments may reveal these misinterpretations, correcting them requires additional time and effort. This, in turn, could be another reason why alleged efficiency differences failed to show up.

This anomaly notwithstanding, participants' hypotheses are generally consistent with Klahr and Dunbar's (1988) conception of scientific discovery as dual search. That is, the overall domain specificity scores from Table 2 confirm expected differences in strategy use based on prior and evolving domain knowledge. However, case-by-case analysis revealed that nearly one-third of the participants acted as Experimenters on the concrete task, and four participants followed an alternative approach not observed in the Klahr and Dunbar (1988) study. A difference in classification methods may be one reason why these participants did not act according to expectations. Klahr and Dunbar (1988) utilized a subset of eight so-called common hypotheses (accounting for about 50% of the experiments) to characterize their participants as Theorists or Experimenters. Classification in the present study occurred on the basis of all of a participant's hypotheses. This all-inclusive method may have revealed more subtle deviations from the Theorists

and Experimenters approach that otherwise might have remained hidden, thus giving a more detailed account of participants' strategy use.

The scope of the inquiry tasks could be another reason why some participants did not follow one of the approaches predicted from the extended SDDS model. Klahr and Dunbar (1988) asked participants to discover the effect of a single programming button on the behavior of a computer-controlled robot tank. Participants in the present study had to discover the effect of four factors, two of which were interrelated. The presence of multiple and interacting factors makes a task less straightforward, and may confuse participants to such an extent that they start floundering (De Jong & Van Joolingen, 1998) or revert from a theory-driven to a data-driven approach – as was observed in 11 participants on the concrete task. This issue merits further attention because learners who are puzzled by the patterns of regularities in a domain may benefit in particular from a systematic, goal-oriented mode of inquiry.

The reciprocal influence of domain knowledge on strategy use could not be assessed due to disproportional strategy use on the abstract task. Only five participants followed an Experimenters' approach to this task (as opposed to 23 Theorists). A comparison of performance measures between Theorists and Experimenters (and possibly participants who adopted the ''other'' approach) would thus result in cell frequencies too low for meaningful statistical analysis. Yet, these comparisons are important to understand how evolving knowledge depends on existing knowledge and how this relation is mediated by investigative strategy use. Future research should therefore compare the Theorist and Experimenter approach on performance success and efficiency, and establish whether observed effects differ as a function of a participant's prior knowledge. By repeatedly assessing knowledge gains and recording participants' hypotheses throughout the inquiry, this research might help uncover how knowledge elements shift from the universal and/or learner hypothesis space to the learner domain space. These comparisons might become even more interesting by manipulating in-between students' levels of prior knowledge, for this would provide additional evidence on the validity of the extended SDDS model.

Overall, the results of this study point to implications for practice. However, the relative simple nature of the experimental tasks could challenge the generalizability of this study's findings to inquiry learning in science classrooms. Admittedly, these tasks do not reflect all attributes of authentic scientific reasoning, but they do call for hypothesis formation and experimentation, which are generally considered key processes in scientific inquiry (Wilhelm & Beishuizen, 2003). More importantly, the current tasks compare favorably to the inquiry tasks found in many textbook-based science curricula, which require students to evaluate the effects of a single independent variable on a single dependent variable (Chinn & Malhotra, 2002). So, although this study does not capture all aspects of true scientific reasoning, it does represent, and hence, generalize to the prevailing hands-on research activities conducted in science classrooms.

Practical implications pertain to offering domain information during simulation-based inquiry learning. This support should facilitate learners in utilizing a theory-driven approach since this mode of inquiry is prevalent as well as consistent with the scientific method: a prescriptive model of research activities that is integral to scientific inquiry learning in many schools. The content and presentation of domain information then depends on the learners' familiarity with and approach to the task at hand. Learners with little prior domain knowledge in general, and the ones that do not follow a Theorists' approach in particular, might benefit from domain information before they start working with the simulation. This would fill their otherwise empty learner domain space and thus enable learners to engage in a theory-driven mode of experimentation. Leutner (1993) corroborated that presenting domain information prior to the students' inquiries is beneficial to learning. However, presenting this information in conjunction with the simulation appeared to be even more effective (cf. Hulshof & De Jong, 2006). This could either mean that the students did not truly incorporate the domain information into their learner domain space, or found it difficult to make sense of this information without being exposed to the learning situation it pertains to. These problems might be overcome by introducing relevant pieces of domain knowledge before learners start their investigations and keeping this information available during the inquiry learning session.

Domain support in familiar domains seems of lesser importance because the learners' prior knowledge provides a comparatively well-equipped learner domain space that enables learners to generate and test specific hypotheses. Still, domain support seems pertinent to learners who fall back on a data-driven mode of inquiry once they have examined all hypotheses they could generate from their prior knowledge. As these learners are unable to infer knowledge from the simulation, presenting this information could assist them in generating and testing new hypotheses. The exact timing of domain information presentation may vary from learner to learner and is difficult to anticipate: one never knows when learners cannot formulate a new hypothesis. Domain support in familiar domains should therefore

be available on demand throughout the entire learning process. Another, more challenging possibility would be to present domain information in adaptive response to the learners' information needs.

A potential threat to the effectiveness of domain support is its interaction with other forms of learner support. Zohar and Peled (2008) showed that explicit metacognitive strategy instruction promotes inquiry learning performance and outcomes. Manlove, Lazonder, and De Jong (2007) further showed that domain support is consulted less often and is less effective when offered concurrently with metacognitive support. These findings not only point at the importance of metacognitive skillfulness in inquiry learning, but also suggest that the presence of multiple and sometimes related support structures can be overwhelming and may add to the learners' cognitive load. A careful orchestration of learner support measures is called for; yet researchers are still trying to find out how different types of support can best be attuned (e.g., Kester, Kirschner, & Van Merriënboer, 2006). Continuing and possibly extending these efforts seems important in view of the fact that inquiry learning is increasingly being mediated by technology-enhanced environments where students regulate their own learning and support is embedded within the environment.

Acknowledgments

The authors gratefully acknowledge the assistance of Emiel van Lieburg in collecting and scoring the data. Ton de Jong is acknowledged for his insightful comments provided during the preparation of this paper.

References

Alexander, P. A., & Judy, J. E. (1988). The interaction of domain-specific and strategic knowledge in academic performance. *Review of Educational Research, 58*, 375–404.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2002). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: a theoretical framework for evaluating inquiry tasks. *Science Education, 86*, 175–218.

De Jong, T. (2006a). Computer simulations: technological advances in inquiry learning. *Science, 312*, 532–533.

De Jong, T. (2006b). Scaffolds for scientific discovery learning. In J. Elen, & R. E. Clark (Eds.), *Handling complexity in learning environments: Theory and research* (pp. 107–128). London: Elsevier.

De Jong, T., & Van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research, 68*, 179–202.

Gijlers, H., & De Jong, T. (2005). The relation between prior knowledge and students' collaborative discovery learning processes. *Journal of Research in Science Teaching, 42*, 264–282.

Hmelo, C. E., Nagarajan, A., & Roger, S. (2000). Effects of high and low prior knowledge on construction of a joint problem space. *Journal of Experimental Education, 69*, 36–56.

Hulshof, C. D., & De Jong, T. (2006). Using just-in-time information to support discovery learning about geometrical optics in a computer-based simulation. *Interactive Learning Environments, 14*, 79–94.

Hulshof, C. D., Wilhelm, P., Beishuizen, J., & Van Rijn, H. (2005). FILE: a tool for the study of inquiry learning. *Computers in Human Behavior, 21*, 945–956.

Ishihara, S. (1982). *Ishihara's test for colour deficiency*. Tokyo: Kanehara.

Kester, L., Kirschner, P. A., & Van Merriënboer, J. J. G. (2006). Just-in-time information presentation: improving learning a troubleshooting skill. *Contemporary Educational Psychology, 31*, 167–185.

Klahr, D., & Carver, S. M. (1995). Commentary: scientific thinking about scientific thinking. *Monographs of the Society for Research in Child Development, 60*(4, Serial No. 245).

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

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: effects of direct instruction and discovery learning. *Psychological Science, 15*, 661–667.

Lavoie, R. D., & Good, R. (1988). The nature and use of prediction skills in a biological computer simulation. *Journal of Research in Science Teaching, 25*, 335–360.

Leutner, D. (1993). Guided discovery learning with computer-based simulation games: effects of adaptive and non-adaptive instructional support. *Learning and Instruction, 3*, 113–132.

Manlove, S., Lazonder, A. W., & De Jong, T. (2007). Software scaffolds to promote regulation during scientific inquiry learning. *Metacognition and Learning, 2*, 141–155.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist, 59*, 14–19.

Schauble, L., Glaser, R., Raghavan, K., & Reiner, M. (1991). Causal models and experimentation strategies in scientific reasoning. *The Journal of the Learning Sciences, 1*, 201–238.

Swaak, J., & De Jong, T. (1996). Measuring intuitive knowledge in science: the development of the what-if test. *Studies in Educational Evaluation, 22*, 341–362.

Tuovinen, J. E., & Sweller, J. (1999). A comparison of cognitive load associated with discovery learning and worked examples. *Journal of Educational Psychology, 91*, 334–341.

Van Joolingen, W. R., & De Jong, T. (1997). An extended dual search space model of learning with computer simulations. *Instructional Science, 25*, 307–346.

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

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

Wilhelm, P., & Beishuizen, J. J. (2004). Asking questions during self-directed inductive learning: effects on learning outcome and learning processes. *Interactive Learning Environments, 12*, 251–264.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*, 172–223.

Zohar, A., & Peled, B. (2008). The effects of explicit teaching of metastrategic knowledge on low- and high-achieving students. *Learning and Instruction, 18*, 337–353.