Instr Sci (2009) 37:437–451
DOI 10.1007/s11251-008-9055-8

Unraveling the influence of domain knowledge during simulation-based inquiry learning

Ard W. Lazonder · Pascal Wilhelm · Emiel van Lieburg

Received: 7 December 2007 / Accepted: 8 April 2008 / Published online: 18 April 2008

Abstract This study investigated whether the mere knowledge of the meaning of variables can facilitate inquiry learning processes and outcomes. Fifty-seven college freshmen were randomly allocated to one of three inquiry tasks. The concrete task had familiar variables from which hypotheses about their underlying relations could be inferred. The intermediate task used familiar variables that did not invoke underlying relations, whereas the abstract task contained unfamiliar variables that did not allow for inference of hypotheses about relations. Results showed that concrete participants performed more successfully and efficiently than intermediate participants, who in turn were equally successful and efficient as abstract participants. From these findings it was concluded that students learning by inquiry benefit little from knowledge of the meaning of variables per se. Some additional understanding of the way these variables are interrelated seems required to enhance inquiry learning processes and outcomes.

Keywords Inquiry learning · Prior knowledge · Induction · Deduction · Computer simulations

Contemporary notions of learning emphasize that learners should be active agents of their own learning processes (Bransford et al. 2002). These educational beliefs coincide with the tenets of inquiry learning, a pedagogy in which learners infer knowledge about relationships between variables in a domain by generating hypotheses and designing and executing experiments to validate these hypotheses. Although inquiry learning is generally assumed to lead to a deeper and more meaningful understanding, the effectiveness of this mode of

A. W. Lazonder (✉) · E. van Lieburg
Department of Instructional Technology, University of Twente, Behavioral Sciences, P.O. Box 217, 7500 AE Enschede, The Netherlands
e-mail: a.w.lazonder@utwente.nl

P. Wilhelm
School of Applied Psychology and Human Resource Management, Saxion Universities of Applied Sciences, P.O. Box 510, 7400 AM Deventer, The Netherlands

learning is challenged by intrinsic problems many learners have with the complex, integrated set of skills inquiry learning entails (De Jong and Van Joolingen 1998).

These difficulties might at least in part be attributable to the learners' low levels of domain knowledge. The basic principle of inquiry learning is for learners to infer knowledge about a task or domain they are relatively unfamiliar with. Research in the area of problem solving has shown that learners with little prior knowledge select less sophisticated strategies and execute these strategies less effectively than learners with higher levels of prior knowledge (e.g., Alexander and Judy 1988). Other studies have shown that this conclusion generalizes to inquiry learning (Hmelo et al. 2000; Lavoie and Good 1988; Schauble et al. 1991).

Offering domain information is an effective means to compensate for low levels of prior knowledge. Leutner (1993) for instance found that presenting domain information prior to the students' inquiries is beneficial to learning. Presenting this information during the inquiry session appeared to be even more effective. Hulshof and De Jong (2006) combined domain support with experimentation hints. Students who received this support during their inquiry achieved higher learning gains than students from the control condition who had no access to this information. Low prior knowledge students in particular benefited from the supportive content they could consult during task performance.

Despite these beneficial effects, offering domain information seems somewhat incompatible with the basic tenets of inquiry learning. In order to maintain the inquisitory nature of the learning process, learners should be given just enough information to overcome learning obstacles, thus enabling them to discover the largest possible part of a domain as efficiently and effectively as possible. Researchers and practitioners should therefore try to strike the right balance between offering and discovering domain information. Although this choice will to some extent depend upon the characteristics of the learner and domain at hand, a general understanding of how domain knowledge shapes the inquiry learning process might certainly be helpful to differentiate the information that should be presented from the information that should be left undisclosed.

Cognitive psychologists use the term content effects to refer to the influence of a person's familiarity with the information presented in the task on task performance. Content effects have been investigated quite extensively in studies of inductive and deductive reasoning. Proficient inquiry learners apply both methods of reasoning: they induce hypotheses from prior knowledge or observed regularities in the domain, and deduce the validity of these hypotheses from evidence gathered through experimentation (Wilhelm and Beishuizen 2003). The results of studies on deductive and inductive reasoning could thus help shed more light on the domain characteristics that facilitate inquiry learning processes and outcomes.

Content effects in deductive reasoning

Early studies on deductive reasoning focused on syllogisms: a kind of logical argument that contains two premises that define a relationship between two categories (e.g., "All cows are animals", and "All cows give milk") and a conclusion that can be inferred from these premises ("Therefore, some animals give milk"). According to Henle's (1962) formal rule theory, premises are often open to different interpretations, which can lead to reasoning errors. This was substantiated by Ceraso and Provitera (1971), who found that people tend to misinterpret premises in abstract syllogisms and, as a result, accept illogical conclusions. Revlis (1975) demonstrated that the use of meaningful content (as in the

above example) can make premises less ambiguous and therefore less susceptible to misinterpretation.

Content effects in deductive reasoning have also been widely studied on Wason's Selection Task (Wason 1966). The initial experiment clearly demonstrated the challenging nature of this seemingly simple task: a mere 27 of the 128 participants managed to solve it correctly (Johnson-Laird and Wason 1970). In a follow-up study, Wason and Shapiro (1971) showed that task performance could be improved considerably if the original abstract task content was replaced by a concrete content. A similar content effect was found by Johnson-Laird et al. (1972) who used a version of the selection task in which participants had to assume the role of postal workers checking the amount of postage on different types of envelopes.

However, this content effect failed to show in a replication study with American participants who were unfamiliar with the postal regulations underlying the task (Griggs and Cox 1982). This could mean that in addition to concrete content, knowledge of the rules that govern the relationships between variables is needed to facilitate deductive reasoning. This was corroborated by Cheng and Holyoak (1985), who found that content effects on the postal task can occur when participants are given a rationale for the rule. Cheng and Holyoak (1985) also examined a version of the selection task in which the variables made only a broad reference to domain-specific content ("If one is to take action *A*, then one must first satisfy precondition *P*"). This task was solved significantly more often than the abstract selection task (61 against 19% of the participants). This finding suggests that, although the precise nature of action *A* and precaution *P* was unknown, participants' knowledge of the causal relationship between actions and precautions in general might have facilitated deductive reasoning.

Overall these studies lend support to the notion that domain knowledge can have a facilitative effect on deductive reasoning. This effect comes about if the rule contains concrete variables that are part of the learner's prior knowledge so that the meaning of the variables evokes the rationale for the rule. This evocation effect even occurs in case of semi-concrete or partially familiar variables. Elucidating the rationale for the rule can be an effective means to facilitate deductive reasoning in cases where the variables make little or no reference to the learner's prior knowledge.

Content effects in inductive reasoning

In inductive reasoning, a person seeks to infer a general rule from a set of specific instances. Presenting the rationale for that rule is inappropriate to facilitate inductive reasoning, for this would conflict with the nature of the reasoning task. Using meaningful content seems a more suitable option, and has been the focus of inductive reasoning research.

Bruner et al. (1956) were among the first to investigate content effects in inductive reasoning. One of their studies used a concrete and an abstract concept attainment task that asked participants to discover a rule that described certain attributes of the concept. In order to figure out the rule, participants could select a presumed instance of the concept and ask the experimenter whether it satisfied the rule. The outcomes of this study indicated that participants in the concrete condition selected more instances and stated more incorrect hypotheses than participants in the abstract condition. From these findings Bruner et al. (1956) concluded that participants in the concrete condition had relied upon the guidance of common experience in attempting to attain the concept, which caused them to perform less efficiently than participants from the abstract condition.

Klahr and Dunbar (1988) reached a different conclusion. Participants in their study had to discover how an unknown "repeat key" influenced the behavior of a computer-controlled robot tank. Although all participants completed the task successfully, the way participants arrived at their solution appeared to depend on their familiarity with the domain. Participants with high prior knowledge required fewer trials to discover the rule that determined the effect of the mystery button. They also performed fewer exploratory trials that were not guided by a hypothesis than their less knowledgeable counterparts. These results indicate that prior domain knowledge has a facilitative effect on performance efficiency. Further analyses revealed that this effect arose because high prior knowledge participants formulated stronger initial hypotheses and generated subsequent, alternative hypotheses more easily.

Similar effects were reported by Wilhelm and Beishuizen (2003), who compared performance across an inductive reasoning task with an everyday-life, concrete content and an isomorphic, abstract task that involved variables and relations participants were incognizant of. Performance success scores differed in favor of participants from the concrete condition. Students who performed the concrete task also stated more experimentation plans, proposed more hypotheses, and made more inferences than students from the abstract condition. Concerning the difference in number of hypotheses, Wilhelm and Beishuizen (2003) postulated that the everyday nature of the concrete task might have invoked expectations about the relevance of independent variables and their effect on the dependent variable. A related explanation is that familiar material provides explanatory mechanisms participants in the concrete condition could use to readily interpret obtained data and induce new hypotheses, thus reducing the number of exploratory trials needed to generate a hypothesis.

Initial evidence for these explanations was acquired by Lazonder et al. (in press). Their study involved a within-subject comparison of performance on a concrete and abstract task. As expected, participants in this study attained higher performance success scores on the concrete task. Their performance on the concrete task was also less exploratory in that they conducted more experiments that aimed to test a hypothesis. A closer inspection of these hypotheses indicated that all participants immediately started formulating and testing hypotheses on the concrete task. On the abstract task, they generally performed four exploratory trials before stating their first hypothesis.

To conclude, the use of concrete, familiar material can facilitate inductive reasoning. When people know the meaning of the instances or variables in the reasoning task, they are more efficient and successful in inducing the rule that governs the relationships between the variables. This facilitative effect is due to the fact that concrete materials have the power of evoking hypotheses about the to-be-discovered rule. Another, related reason is that concrete materials provide explanatory mechanisms which facilitate the ease with which new, alternative hypotheses can be inferred from obtained data.

Investigating content effects in inquiry learning

Inquiry learning revolves around discovering relationships between variables in an unfamiliar domain. Toward this end learners have to induce hypotheses from prior knowledge or data obtained through exploration, and deduce the validity of these hypotheses through systematic experimentation. At the outset of this paper it was assumed that generating and testing hypotheses proceeds more efficiently if learners possess knowledge of the task or

domain at hand. Reasoning research suggests that knowledge of the meaning of variables can facilitate the discovery of rules that govern the relationships between these variables.

However, the conclusions from reasoning research do not necessarily generalize to inquiry learning. The inductive reasoning studies in particular used materials from everyday-life in which knowledge or expectations about the relationships was implied by the meaning of the variables. In Wilhelm and Beishuizen (2003) for instance, participants had to discover how common factors such as type of bicycle or weight of a schoolbag influence the time it takes a teenage boy to cycle to school. As virtually everybody has at least some sense of the general distinction between a city bike and a racing bike, people will automatically assume the latter reduces the boy's traveling time. Since inquiry learning is essentially about discovering relations between variables in unfamiliar domains, it is unclear whether the same inferential mechanism will apply. To illustrate, high school students who know what current, potential difference, and resistance mean, may not readily see how the former variable is determined by the latter two in electrical circuits containing a conductor.

An experiment was performed to investigate whether knowledge of variables alone is sufficient to facilitate inquiry learning in unfamiliar domains. The main purpose of this experiment was to establish whether inquiry learning processes and outcomes are enhanced in cases where the relation between variables is neither implied by nor inducable from the meaning of the variables. Toward this end the experiment compared performance across three isomorphic inquiry tasks. The concrete task involved a realistic problem and included variables that had a familiar meaning to all participants. The relations between the variables could readily be inferred from the meaning of the variables. The intermediate task also contained variables that are familiar from everyday life, but unfamiliar in relation to the other variables in general and the dependent variable in particular. As a result, relationships could not be inferred from the meaning of the variables. The abstract task was designed so that participants would neither know the meaning of the variables from everyday life, not be able to infer expectations about the relations from the meaning of variables.

Two sets of pairwise comparisons were made to establish whether knowledge of the meaning of variables is a sufficient condition to facilitate inquiry learning. If so, participants in the intermediate condition would have to perform as efficiently and successfully as participants in the concrete condition *and* more efficiently and successfully than participants in the abstract condition. If not, concrete participants would outperform intermediate participants in terms of performance efficiency and performance success whereas the comparison across the intermediate and abstract condition would show no differences on these measures.

Method

Participants

Fifty-seven first-year students in social sciences volunteered to participate in the experiment for course credits. There were 26 males and 31 females with a mean age of 19.61 years ($SD = 1.79$). Forty-three participants were Dutch; fourteen 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. Participants were randomly assigned to experimental conditions, leading to 19 participants per condition.

Inquiry tasks

Three inquiry tasks were created in an authoring environment named FILE (Hulshof et al. 2005). The concrete task invited participants to investigate how each of four factors affects 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 substance 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 10k 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 inhibited 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.

Factors in the concrete task are generally well-known, and were thus expected to be familiar to the college freshmen who participated in the experiment. Their knowledge of the difference between factor values was further assumed to evoke expectations about the factors' general direction of effect on the 10k time. In case of nutrition for instance, participants would infer from their knowledge of the difference between the three types of food that sport food yields superior performance compared to junk food. Inspired by the characteristics of the tasks of Klahr and Dunbar (1988) and Lazonder et al. (in press), the factor Xelam was slightly ambiguous (different substances have different effects, not all of which are beneficial to track running). This served to maintain the inquisitory nature of the concrete task—participants might display atypical inquiry behavior or could not take the task seriously if it merely contained known variables and inducable relationships. A pilot test confirmed that college freshmen generally believe Xelam is some kind of drug that improves performance during a 10k race, but need to scrutinize its effects in the simulation to validate their presumption.

The intermediate task contained four familiar factors, namely musical instrument (drum, guitar, or trumpet) cutlery (fork, spoon, or knife), flower (tulip or daffodil), and clothing (trousers or shirt). Combinations of selected values for each factor yielded a number of points, and participants had to discover how each factor influenced this score. Inference of relations could not be determined by the meaning of the factors. It was therefore highly implausible (if not impossible) that participants' prior knowledge of the difference between a factor's values would evoke expectations about its influence on the numerical value. The task structure defining the influence of the factors was made similar to that of the concrete task by substituting factors. Musical instrument replaced nutrition, cutlery took the place of training frequency, flower supplanted smoking, and clothing substituted Xelam. Factor values were transformed accordingly (e.g., continue smoking = tulip; quit smoking = daffodil).

The abstract task was designed to ensure that participants would have no prior knowledge of the difference between the factor values, and could not infer expectations about the relations from the factors' meaning. This task asked participants to discover how four geometrical shapes influenced 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 as follows: training frequency was replaced by the circle, smoking became the triangle, nutrition was substituted by the square, and Xelam by the question mark. The values of the factors were adapted accordingly.

The operation of the simulation was identical for all tasks. Participants could discover the impact of a single factor by manipulating its values and observing the effect on the output variable (either 10k time or number of points). Factor values could be set by clicking the corresponding icon on the left side of the screen (see Fig. 1). 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.

Instruments

A background questionnaire determined participants' demographic characteristics. A color vision deficiency test was administered to ensure that participants in the abstract condition were able to perceive differences between the colored shapes in their version of the task. (As factor values in the concrete and intermediate task had other distinguishing features besides color, normal color vision was not a requirement in these conditions.) The test utilized seven items from the Ishihara color test (Ishihara 1982) that screened 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

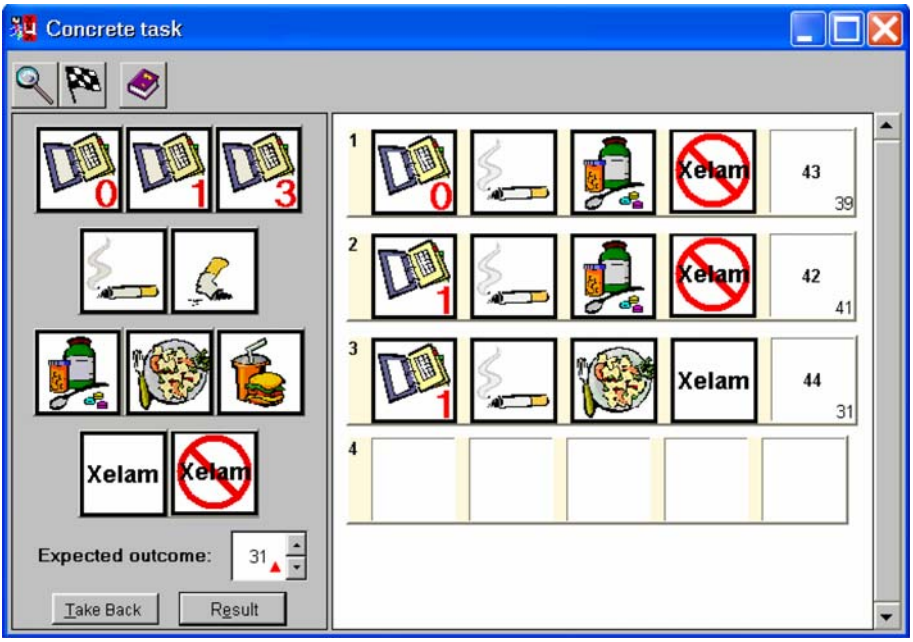

Fig. 1 Simulation interface of the concrete task

appropriate for screening purposes. The K-R 20 reliability estimate of the seven items was .95.

A pretest was administered to check whether participants could make the correct inferences about the factors' general direction of effect. (The magnitude of these effects on the performance of the athlete in this task had to be discovered during the experiment by interacting with the simulation.) The ten items addressing training frequency, smoking, and nutrition were "what-if" questions (Swaak and De Jong 1996). Each item comprised an initial situation (e.g., "During an 8-week preparation for a 10k 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 one post-change situation. Two additional items assessed participants' conceptions of Xelam and its effect on the 10k time. As the meaning of Xelam was slightly ambiguous, and its effects on the 10k time depended on training frequency, the what-if format with preset alternatives would reveal rather than assess the differentiated relationship participants should infer from the meaning of this factor. Hence two open questions 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 what-if items addressed the effect of a 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 by interacting with the simulation—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 = .61).

Procedure

Students participated in the experiment one at a time, receiving the same instructions and following the same experimental procedures. All sessions took place in a quiet room equipped with one computer. 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). These tests captured student characteristics relevant to the concrete and abstract task, respectively, but were administered in all conditions to rule out internal validity threats. Although the pretest could trigger inferences about the relationships in the concrete task, this was not deemed undesirable since being able to make these inferences was a requirement to participants in the concrete condition. For the same reason, participants in the intermediate and abstract task (who should *not* be able to make inferences prior to the task) received the same pretest and were *not* asked about the variables and relations in their version of the task.

Following test administration 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 either the concrete, intermediate, or abstract task, depending on the condition they were assigned to.

After having read the task's problem statement, they 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. The maximum time for task completion was 40 min.

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 and Carver 1995), research has shown that non-directive probes have no influence on participants' inquiry learning processes (Wilhelm and 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 satisfactory inter-rater agreement (90% agreement).

Coding and scoring

Participants' responses to the seven items of the color vision deficiency test were scored as true or false. Their responses to the pretest were checked against the model underlying the concrete task. One point was allocated to each response that matched the relationship between the variables in this model. Two open questions assessed participants' conceptions of Xelam. Their responses were coded according to the presumed direction of effect (i.e., positive effect, negative effect, no effect). 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.

Main variables of the study were performance efficiency and performance success. Performance efficiency was determined by time on task, the experiments participants conducted, and the hypotheses they proposed. Time and experiments were assessed from the logfiles. For the latter measure, a distinction was made between the number of unique and duplicated experiments. Simulation runs testing a new, untried combination of factor values were coded as unique experiment; if this combination reappeared in subsequent trials it was considered a duplicated experiment. The maximum number of unique experiments was 36 since there were this many distinct combinations of factor values in each task. Each experiment was further classified according to the presence or absence of a hypothesis (the definition of a hypothesis is given below). This categorization was used to calculate the percentage of exploratory experiments as the ratio of the number of experiments that were not guided by a hypothesis to the total number of experiments.

Participants' hypotheses were scored from their responses to the probing questions. A hierarchical rubric was used to classify the 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. Inter-rater agreement scores (Cohen's κ) were .97 for the concrete task, .96 for the intermediate task, and .97 for the abstract task.

Performance success was 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. Inter-rater reliability estimates (Cohen's κ) reached .86 for the concrete task, .92 for the intermediate task, and 1.00 for the abstract task.

Results

The color vision deficiency test checked whether participants in the abstract condition were able to visually discriminate factor values. One male participant from the abstract condition answered one item incorrectly; all other participants achieved a perfect score. These scores gave no reason to assume that participants in the abstract condition had difficulties distinguishing factor values. Pretest scores indicated whether participants in the concrete condition could infer the nature of the relationships between the variables in the simulation. Participants in the concrete condition produced a mean pretest score of 9.16 ($SD = 1.12$). However, one participant achieved a pretest score of 6. As this score was more than two standard deviations below the sample mean, this participant was excluded from the sample. The remaining participants' conception of Xelam was quite consistent: 17 participants thought it would generally improve the athlete's performance; two participants thought Xelam would have no effect.

Table 1 summarizes the descriptive statistics for performance efficiency. Data for time on task indicate that participants generally needed less than the given 40 min to complete their task. Univariate analysis of variance (ANOVA) showed that task type had no effect on this measure ($F(2,53) = 1.93$, $MSE = 123.19$, $p = .16$), indicating that participants in the three conditions needed as much time to complete their version of the task.

Multivariate analysis of variance (MANOVA) yielded no significant effect of task type on the number of unique and duplicated experiments ($F(4,106) = 1.25$, $p = .29$). There was, however, a significant difference in exploratory experiments ($F(2,53) = 11.28$, $MSE = 549.14$, $p < .01$). Planned contrasts revealed that participants in the concrete condition had a significantly lower percentage of exploratory experiments than participants from the intermediate condition ($t(53) = 4.45$, $p < .01$, $r = .52$). The difference in

Table 1 Means and standard deviations for performance efficiency by type of task

	Type of task		
	Concrete ($n = 18$)	Intermediate ($n = 19$)	Abstract ($n = 19$)
Time on task			
Time (min)	28.76 (11.62)	25.50 (12.31)	32.56 (9.14)
Experiments			
Unique experiments (#)[a]	15.56 (5.92)	14.16 (6.21)	18.16 (5.53)
Duplicated experiments (#)	5.56 (8.76)	5.0 (5.61)	5.95 (5.68)
Exploratory experiments (%)	12.27 (5.52)	46.58 (5.38)	40.86 (5.38)
Hypotheses[b]			
Unspecified hypotheses (%)	5.68 (6.86)	4.15 (8.94)	10.84 (18.02)
Partially-specified hypotheses (%)	80.07 (16.23)	60.36 (37.70)	34.52 (36.97)
Fully-specified hypotheses (%)	14.25 (16.99)	35.50 (38.13)	54.64 (32.58)

[a] Maximum = 36

[b] Scores for the three types of hypotheses within each condition add up to 100%. One intermediate participant generated no hypotheses and was removed from the analyses

percentage of exploratory experiments between the intermediate and abstract condition was not significant ($t(53) = .75$, $p = .46$).

A MANOVA on the participants' hypotheses produced a significant effect of task type ($F(4,104) = 4.40$, $p < .01$). Subsequent univariate ANOVAs showed that task type had no effect on the percentage unspecified hypotheses ($F(2,52) = 1.49$, $MSE = 228.95$, $p = .24$), but did affect the percentage of partially-specified and fully-specified hypotheses ($F(2,52) = 9.44$, $MSE = 9670.28$, $p < .01$; respectively $F(2,52) = 8.05$, $MSE = 7541.14$, $p < .01$). Planned contrasts revealed that participants in the concrete and intermediate condition had comparable percentages of partially-specified hypotheses ($t(52) = 1.85$, $p = .07$), whereas intermediate participants had a lower percentage of partially-specified hypotheses than participants in the abstract condition ($t(52) = 2.45$, $p < .05$, $r = .32$). For fully-specified hypotheses this pattern in scores was reversed, with a lower percentage of hypotheses in the concrete condition compared to the intermediate condition ($t(52) = 2.08$, $p < .05$, $r = .28$), and a nonsignificant difference among the intermediate and abstract condition ($t(52) = 1.90$, $p = .06$).

A performance success score reflected the extent to which participants' knowledge of the relations between the variables matched the simulations' underlying model. An ANOVA produced a significant effect of task type on this measure ($F(2,53) = 3.25$, $MSE = 7.54$, $p < .05$). Planned contrasts showed that this effect arose because participants in the concrete condition ($M = 9.56$, $SD = 2.38$) achieved significantly higher performance success scores than participants in the intermediate condition ($M = 7.37$, $SD = 3.08$; $t(53) = 2.42$, $p < .05$, $r = .32$). Scores in the intermediate condition were lower than those in the abstract condition ($M = 9.05$, $SD = 2.72$), but this difference was not statistically significant ($t(53) = 1.89$, $p = .06$).

Multiple regression analysis was performed to explain observed differences in performance success from the measures that defined performance efficiency. Using the Enter method, a significant model emerged ($F(7,48) = 4.10$, $MSE = 5.85$, $p < .01$) that accounted for 28.3% of the variance in performance success. As the unstandardized regression coefficients in Table 2 indicate, performance success was negatively related to

Table 2 Summary of regression analysis for variables predicting performance success

Predictor variable	*B*	*SE*	β
Time on task	−.032	.048	−.125
Unique experiments	−.038	.102	−.080
Duplicated experiments	.081	.060	.190
Exploratory experiments	−.047	.014	−.450**
Unspecified hypotheses	.078	.036	.339*
Partially-specified hypotheses	.055	.027	.710*
Fully-specified hypotheses	.073	.026	.876**

Note: Adjusted $R^2 = .283$ ($N = 56$, $p < .01$)

* $p < .05$

** $p < .01$

the percentage of exploratory experiments. With all other variables held constant, a 10% decrease in exploratory experiments would increase the performance success score by nearly half a point. Performance success was positively associated with the participants' hypotheses, meaning that a 10% increase in unspecified hypotheses would raise performance success scores by .78. For partially specified and fully specified hypotheses, these values were .55 and .73, respectively. The other variables listed in Table 2 did not make a significant contribution to the regression model.

Together these findings suggest that the mere testing of hypotheses, regardless of their level of domain specificity, was the strongest predictor of performance success. Stepwise multiple regression analysis bore this out. Using the significant predictor variables above, exploratory experiments was the only parameter in the model that had a significant impact on performance success scores (Adjusted $R^2 = .223$, $\beta = -.487$, $p < .001$).

Discussion

This study sought to uncover why domain knowledge facilitates inquiry learning processes and outcomes. Reasoning research has shown that knowledge of the meaning of variables is sufficient to facilitate the discovery of rules that govern the relationships between these variables. If this conclusion generalizes to inquiry learning, participants in the concrete and intermediate condition would perform equally efficient and successful. If not, performance efficiency and performance success would be comparable across the intermediate and abstract condition.

Results for performance efficiency show that participants in the intermediate and concrete condition needed as much time to complete their task and performed as many experiments. In the concrete condition however, these experiments were more often used to test a hypothesis, as evidenced by the lower percentage of exploratory experiments. This suggests that the concrete task enables participants to infer hypotheses from the meaning of the variables; additional evidence for this claim comes from the pretest scores. Multiple regression analysis further showed that the presence of hypotheses per se is a plausible explanation for the fact that participants in the concrete condition achieved higher performance success scores than participants from the intermediate condition.

In view of these findings one might expect comparable outcomes for participants in the intermediate and abstract condition. Performance efficiency measures support this

expectation: participants in the aforementioned conditions completed their task equally fast, conducted as many experiments, and displayed a similar percentage of exploratory experiments. Performance success was also comparable in these conditions. Although intermediate participants admittedly tended to perform less successful than their abstract counterparts, this difference was not statistically significant. From these pairwise comparisons it can be concluded that students learning by inquiry have little to no advantage of knowledge of the meaning of variables. Some additional understanding of the way these variables are interrelated, or the possibility to infer these relations, seems required to facilitate inquiry learning processes and outcomes.

Still, the intermediate and abstract condition differed with regard to the specificity of the participants' hypotheses. Lazonder et al. (in press) showed that this measure is indicative of participants' approach to the task. On the concrete task, inferential mechanisms enable participants to infer hypotheses about the direction of effects ("the more you practice, the faster you will run"). Testing this partially-specified hypothesis produces outcomes (e.g., a 4-min difference in time) that can readily be interpreted on the basis of prior knowledge, thus reducing the need to validate observed outcomes by generating and testing a fully-specified hypothesis. Participants in the abstract condition—in absense of prior knowledge and inferential mechanisms—start off in a data-driven mode. They perform exploratory experiments to get some sense of the relations between dependent and independent variables ("maybe the blue triangle yields a two-point higher score than the brown triangle") and then generate and test fully-specified hypotheses to assess whether newly induced knowledge is correct.

As hypotheses on the intermediate task hold the middle ranks on specificity, participants in this condition may have combined both approaches by performing exploratory experiments to generate and test fully-specified hypotheses and inferring partially-specified hypotheses from prior knowledge and experimental outcomes. The first part of this supposition is substantiated by the high percentage of exploratory experiments in this condition; support for the second part comes from anecdotal evidence gathered by the experimenters. They noticed that initial outcomes often evoked a rationale for the rule governing the relationship between variables. Many participants thought the numerical score represented money or weight, and used this idea to formulate and test partially-specified hypotheses for the other variables. Since this rationale could not be confirmed (relationships were arbitrary), participants may have developed confusion or misconceptions that caused them to perform less successfully than participants in the other conditions. Unfortunately there was no systematic observation of this usage of inferential mechanisms, but it would certainly be interesting to address this issue in future research.

These unexpected inferences nevertheless strengthen the conclusion that a basic understanding of relations between variables is required to enhance inquiry learning processes and outcomes. On the one hand this knowledge can bootstrap students' inquiry practices in cases where relations in a domain cannot be inferred from the meaning of variables. On the other hand this knowledge can safeguard students from the adverse consequences of their quest for meaning. The experimenters' observations suggest that learners have a strong natural tendency to induce a rationale from discovered relations which in turn evokes hypotheses about undiscovered relations. As these inferences are easily prone to error (especially for domain novices), offering some information on the relations in a domain could warrant the efficiency and success of inquiry learning within that domain.

Still, some people might question the generalizability of this study's findings to inquiry learning in schools. Asking students to explore unknown relationships between known or

even unknown variables is admittedly an uncommon educational practice. This study was certainly not designed to advocate the use of such tasks (and neither do its results), but instead served to further improve our understanding of effective scientific inquiry learning in schools. On the other hand, it is not inconceivable that science educators who adopt an inquiry-based pedagogy put at least some of their students in a position comparable to that of participants in the intermediate or abstract condition. Students who have difficulty understanding the material, or did not prepare themselves for class may perceive a computer simulation of a familiar phenomenon as one in which the variables and/or their relations unknown. How these students can be supported on a just-in-time basis is an important question for future research.

Acknowledgements The authors gratefully acknowledge the assistance of Mieke Hagemans in collecting and scoring the data. Yvonne Mulder is acknowledged for her 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: National Academy Press.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.

Ceraso, J., & Proviter, A. (1971). Sources of error in syllogistic reasoning. *Cognitive Psychology, 2*, 400–410.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17*, 391–416.

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

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology, 73*, 407–420.

Henle, M. (1962). On the relation between logic and thinking. *Psychological Review, 69*, 366–378.

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. 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.

Johnson-Laird, P. N., Legrenzi, P., & Legrenzi, M. S. (1972). Reasoning and a sense of reality. *British Journal of Psychology, 63*, 395–400.

Johnson-Laird, P. N., & Wason, P. C. (1970). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology, 1*, 134–148.

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.

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

Lazonder, A. W., Wilhelm, P., & Hagemans, M. G. (in press). The influence of domain knowledge on strategy use during simulation-based inquiry learning. *Learning and Instruction*. doi: 10.1016 /j.learninstruc.2007.12.001.

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.

Revlis, R. (1975). Two models of syllogistic reasoning: Feature selection and conversion. *Journal of Verbal Learning and Verbal Behavior, 14*, 180–195.

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.

Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology* (pp. 135–151). London: Penguin.

Wason, P. C., & Shapiro, D. A. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology, 23*, 63–71.

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.