Journal : Small 11251 | Dispatch : 31-5-2013 | Pages : 20
Article No. : 9280 | LE | TYPESET
MS Code : | CP | DISK

Instr Sci
DOI 10.1007/s11251-013-9280-7

ORIGINAL RESEARCH

The benefit of being naïve and knowing it: the unfavourable impact of perceived context familiarity on learning in complex problem solving tasks

Jens F. Beckmann · Natassia Goode

Received: 14 May 2012 / Accepted: 20 May 2013

Abstract Previous research has found that embedding a problem into a familiar context does not necessarily confer an advantage over a novel context in the acquisition of new knowledge about a complex, dynamic system. In fact, it has been shown that a semantically familiar context can be detrimental to knowledge acquisition. This has been described as the "semantic effect" (Beckmann, Learning and complex problem solving, Bonn, Holos, 1994). The aim of this study was to test two competing explanations that might account for the semantic effect: goal adoption versus assumptions. Participants were asked to learn about the causal structure of a linear system presented on a computer containing three outputs by changing three inputs through goal free exploration. Across four conditions the level of familiarity was experimentally varied through the use of different variable labels. There was no evidence that goal adoption can account for poor knowledge acquisition under familiar conditions. Rather, it appears that a semantically familiar problem context invites a high number of a priori assumptions regarding the interdependency of system variables. These assumptions tend not to be systematically tested during the knowledge acquisition phase. The lack of systematicity in testing a priori assumptions is the main barrier to the acquisition of new knowledge. The semantic effect is in fact an effect of untested presumptions. Implications for research in problem solving, knowledge acquisition and the design of computer-based learning environments are discussed.

Keywords Complex problem solving · Dynamic systems · Knowledge acquisition · Semanticity · Semantic effect

J. F. Beckmann (✉)
School of Education, Durham University, Leazes Road, Durham DH1 1TA, UK
e-mail: j.beckmann@durham.ac.uk

N. Goode
University of the Sunshine Coast Accident Research, Faculty of Arts and Business,
University of the Sunshine Coast, Sippy Downs, QLD 4556, Australia

Journal : Small 11251 | Dispatch : 31-5-2013 | Pages : 20
Article No. : 9280 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

Introduction

Whilst the influence of prior experience on problem solving and learning has been examined quite extensively in studies on reasoning (e.g. Blessing and Ross 1996; Hesse et al. 1997; Kotovsky and Fallside 1989), few studies have considered its impact in the context of complex, dynamic problems. Complex, dynamic problems differ from the static problems traditionally used in psychological research as they not only change as a result of the decisions made by the problem solver but also change autonomously (Schoppek 2002). Complex dynamic problems have been used to investigate human problem solving in complex contexts (complex problem solving, CPS, e.g. Funke 1992); to study knowledge acquisition processes and learning (Beckmann 1994; Beckmann and Guthke 1995; Guthke et al. 1995; Goode and Beckmann 2010); and to assess problem solving competencies (Greiff and Funke 2009) and intelligence (e.g. Kröner et al. 2005). They are also increasingly utilised in educational contexts such as nursing (McGaghie et al. 2006; Ravert 2002), business and management education (Lainema and Nurmi 2006; Wood et al. 2009), engineering (Chung et al. 2001; Fang et al. 2011), or generically to teach scientific principles (de Freitas and Oliver 2006; Goldstone and Son 2005).

A distinction is often made between "abstract" complex problems and "concrete" or "semantically meaningful" complex problems (e.g. Beckmann 1994; Burns and Vollmeyer 2002; Lazonder et al. 2008, 2009). Concrete or semantically meaningful problems use cover stories and variable labels that refer to familiar systems in the real world. For example, in *LOHHAUSEN* the problem solver is instructed to act as the mayor of a virtual small town dealing with variables labelled such as "living standard of the workforce" and "energy consumption" (Dörner 1987), whilst in *FIRECHIEF* individuals are required to control fire station resources labelled as "helicopters" or "trucks" to stop simulated forest fires spreading (Omodei and Wearing 1995). Abstract problems such as *SINUS* (Funke 1992) and *MACHINE* (Beckmann 1994; Beckmann and Guthke 1995) have cover stories and variable labels that do not refer to any known or previously experienced system. In *SINUS* problem solvers have to deal with an ecosystem on a fictitious planet where artificially labelled creatures live in an unknown dependency from one another. In *MACHINE*, as a further example, problems solvers are asked to discover how "Control A", "B" and "C" influence values in "Display X", "Y" and "Z". The terms "abstract" and "concrete" in this context are in fact rather ambiguous. It could be argued that the difference between systems like *FIRECHIEF* and *MACHINE* does not lie in the concreteness or abstractness of variable labels; a fire station is as abstract or concrete as a machine. Goldstone and Son (2005, p. 72) refer to computerised interactive simulations or microworlds that model real world entities as being "virtually concrete". The labels used in either problem, "concrete" or "abstract", are part of common language, and hence are in this regard semantically meaningful, hence virtually concrete. They are likely to differ, however, with regard to their semanticity, the degree to which the semantic context formed by the variable labels creates some sense of familiarity in the problem solver. Here it is important to note, however, that a sense of familiarity is not necessarily based on actual prior experience. We can safely assume that only a minuscule minority of participants in respective studies were in fact head of a fire fighter unit, entrusted with the governing of a town, or involved in the management of a small, shirt producing, textile company "in real life". The presumed sense of familiarity of these kinds of contexts is more often than not derived from "second hand experience" at best, which in itself constitutes what is referred to as "common or world knowledge". Hence, comparisons between "abstract" and "concrete" problems may in fact be a comparison between varying degrees of semantic richness of variable labels

Journal : Small 11251 Dispatch : 31-5-2013 Pages : 20
Article No. : 9280 LE TYPESET
MS Code : CP DISK

(i.e. semanticity), which in turn, may trigger varying amounts of familiarity in the problem solver. The consequences of these processes, especially their effect on what is learned while dealing with semantically embedded complex, dynamic problems, are the main focus of this paper.

The effect of semanticity

One question in this context is whether concrete or abstract variable labels should be used when studying knowledge acquisition in complex, dynamic environments. In their review, Goldstone and Sakamoto (2003) suggest that the use of variable labels that refer to familiar contexts facilitates the understanding of abstract scientific principles. The sense of familiarity is considered helpful to learners to understand the role that each of the variables plays in a system (i.e. cause or effect). According to Klahr (2000), familiar problem contexts may readily promote the formation of assumptions and hypotheses through “analogical mapping, heuristic search, priming, remindings or conceptual combination” (p. 33). In the process of knowledge acquisition or learning problem solvers must then evaluate the available data to determine whether it confirms or disconfirms these hypotheses. If the task is novel and, due to its abstractness, prior experience does not evoke hypotheses about the underlying structure of the system then knowledge must be induced directly and solely from the data (Klahr 2000). From this perspective it could be theorised that a familiar context in fact facilitates the acquisition of knowledge about the underlying structure of a system.

However, empirical evidence does not unequivocally support this claim. In a series of studies by Beckmann (1994) involving secondary school students (mean age 14.3 years) as well as university students (mean age: 24.1 years) participants were given one of two versions of the same complex, dynamic system. The variables in one version, *CHERRY TREE*, were labelled “Light”, “Water” and “Heat” for the inputs and “Cherries”, “Leaves” and “Beetles” for the outputs, whilst the input variables in another version, *MACHINE*, were “Control A”, “Control B” and “Control C”, the outputs were labelled “Display X”, “Display Y” and “Display Z”. The results clearly showed that problem solvers who worked with the *MACHINE* acquired significantly more knowledge about the system’s underlying causal structure than those who worked with the *CHERRY TREE*. An almost intrudingly obvious explanation for this—at first sight rather counterintuitive—result is that the system structure for *CHERRY TREE* must have been counterfactual to “common world knowledge”. However, this can be easily ruled out by analysing the a priori assumptions that problem solvers were asked to report before they interacted with the system. Problem solvers’ expectations regarding the anticipated interdependency of the variables did neither systematically agree nor disagree with the actual underlying structure of the *CHERRY TREE* system. They in fact represented a balance of correct and incorrect assumptions. Hence, the argument of a counterfactual system, which is in conflict with “common sense”, “world knowledge”, or “prior experience”, appears not to be viable.

In another study, Burns and Vollmeyer (2002) also gave participants structurally isomorphic problems with different set of labels for the system variables. In one version, labels were selected to make the links between the variables as obvious as possible (e.g. “hot water” increased “temperature”), whilst the labels in the alternative version of the same problem were selected to give no helpful information (e.g. “lime” increased “oxygenation”). The results indicate that participants dealing with the “suggestive” problem started off at a significant advantage based on prior knowledge that they were able to

Journal : Small 11251 Dispatch : 31-5-2013 Pages : 20
Article No. : 9280 □ LE □ TYPESET
MS Code : ☑ CP ☑ DISK

utilise. However, they did not ultimately gain more knowledge than those dealing with the "less helpful" variable labels. This is even more surprising if we consider that the context-induced assumptions represent knowledge that exists prior to dealing with the system, thus it is knowledge that does not need to be acquired as a result of direct learning in the situation studied.

Similarly, the results of a recent series of studies conducted by Lazonder et al. (2008, 2009) offer, although unwittingly, compelling evidence that even appropriate assumptions about the underlying structure of a system may be detrimental to exploratory learning and may inhibit the acquisition of new knowledge. Lazonder et al. (2009) gave participants structurally isomorphic complex problems that had either "concrete", "abstract" or "intermediate" labels of the system variables. In the "concrete" version of the problem, the labels were selected to make three out of four links as obvious as possible, relying on common world knowledge or experience (e.g. smoking is expected to be detrimental to running performance). In the "abstract" and "intermediate" versions of the problem the labels did not imply any specific links between the variables, thus all the rules regarding variable interdependency had to be induced directly from data generated through direct interaction with the system. As the variable labels in the "concrete" condition formed a familiar context that allowed learners to derive correct assumptions, learners' prior knowledge in this experimental condition provided them with an initial advantage of six points on the knowledge scale used in this study (ranging from 0 to 12). After the learning phase, learners in this condition gained (i.e. learned) on average 3.56 knowledge points whilst the "intermediate" group gained on average more than twice as much (i.e. 7.37 points). Learners under "abstract" conditions gained even more (i.e. 9.05 points). In comparison to the "concrete" condition, the exploration behaviour of the "abstract" condition was characterised by a significantly higher proportion of fully specified hypotheses being tested, which ultimately led to a high level of performance success (estimated Cohen's $d = 1.62$).[1]

In summary, the empirical evidence from these studies (knowingly or unknowingly by authors) challenges the prevailing position that problem solving and learning is facilitated by contextualisation or concreteness. Empirical evidence rather seems to suggest that novel or abstract problem contexts might be advantageous in regard to knowledge acquisition and learning.

Possible causes for the semantic effect

The aim of this paper is to further elucidate the underlying reasons for what has been referred to as the "semantic effect" (Beckmann 1994, p. 118). The overarching research question therefore is why problem contexts with high levels of semanticity seem to inhibit the acquisition of new knowledge. The viability of two possible explanatory mechanisms will be tested: (1) premature goal adoption and (2) semanticity induced presumptions.

The first proposed explanation for the semantic effect builds upon the assumption that under semantically rich or familiarity inducing conditions, problem solvers may try to control the system to reach context-related goals rather than exploring and testing

[1] Curiously, the authors interpreted these results differently. In comparing the final knowledge score between the three experimental conditions—without considering the a priori differences in knowledge—they erroneously arrived at the conclusion that concrete conditions are advantageous to the acquisition of new knowledge.

Journal : Small 11251 Dispatch : 31-5-2013 Pages : 20
Article No. : 9280 LE TYPESET
MS Code : CP DISK

hypotheses regarding its underlying causal structure. Hesse (1982) found that problem solvers when given a problem presented in a familiar context were more goal-directed, and paid less attention to the structure of the system than those given an abstract, hence novel version of the same problem. From this perspective, a familiar context may encourage problem solvers to pursue goals, rather than explore and acquire knowledge about the system structure. Previous research has shown that participants' awareness of future goals (i.e. certain target values for the output variables of the system) shifts attention away from the acquisition of knowledge. In studies that do not have a dedicated exploration phase or explicit targets, problem solvers tend to set their own performance goals (Funke 1992). Vollmeyer and colleagues found that problem solvers who were cognisant of future goals practiced achieving these goals even though they were instructed to explore and to learn about the system (Vollmeyer et al. 1996, 2002). Schauble et al. (1991)refer to the tendency to produce a desired outcome rather than aiming for understanding the underlying model as "engineering approach" (see also Njoo and de Jong 1993). In sum, system characteristics or modes of instruction that allow for any type of goal adoption (i.e. either self or externally set) may impede knowledge acquisition.

An alternative explanation for the emergence of the semantic effect was proposed by Beckmann (1994). He suggested that problem contexts with high levels of semanticity are likely to induce a priori assumptions. This process might be mediated by a sense of familiarity in the problem solver. The acquisition of an accurate representation (i.e. mental model) of the system's actual causal structure would then require a systematic testing of these assumptions. The focus of such systematic testing would need to be on ruling out inappropriate assumptions regarding the interdependence of system variables. Beckmann (1994) further assumed that this *reduction of complexity* imposes higher cognitive demands than inferring individual rules directly from the data (i.e. observed system behaviour) under conditions with low semanticity. The latter represents a process of *construction of complexity*. An evasion of the more demanding process of complexity reduction appears to result in the general tendency to seek information that confirms rather than potentially disconfirms initial assumptions (Dunbar 1993; Klayman and Ha 1987; Wason 1966).

In this study we will test hypotheses derived from both explanations for the origin of the semantic effect. Testing the viability of the *Goal adoption explanation* requires contrasting performances obtained in versions of the same system that differ with regard to their control worthiness[2]; testing the viability of the *Presumption explanation* requires contrasting performances obtained in versions of the same system that differ with regard to the semanticity carried by their variable labels.

Methods

Participants

A convenience sample of 80 first year psychology students at the University of Sydney participated for course credit. Participants were randomly allocated to one of four

[2] We refer to control worthiness as a characteristic of a complex, dynamic system that is determined by the semanticity of its output variables. The underlying assumption is that output variables high in semanticity (i.e. with semantic reference to concrete objects in the "real world") are more likely to trigger control behaviour that aims at optimising levels of output variables according to self-set targets (e.g. increase, decrease, or keep stable) despite the task being to explore the system.

Journal : Small 11251 | Dispatch : 31-5-2013 | Pages : 20
Article No. : 9280 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

conditions, which will be explained in detail in the following section, and were tested in groups of 2–10 participants. Their ages ranged from 18 to 48 years ($M = 20.19$, $SD = 5.19$).

Materials

The underlying causal structure of the complex, dynamic system employed in this study was identical to the one introduced by Beckmann (1994). The system consists of three input variables and three output variables. Figure 1 depicts the causal structure of the system with *CHERRY TREE* labels. The underlying causal structure (i.e. 6 relationships out of 12 possible) remained identical across all versions of the system. The variable labels were systematically varied in order to create four experimental conditions that allowed testing of the hypotheses.

Design

A simple contrast between performances obtained while working with a system with high semanticity (e.g. *CHERRY TREE*) and low semanticity (e.g. *MACHINE*), as done in previous studies, would not suffice to help deciding whether goal adoption or presumptions are causal to the semantic effect. In order to experimentally test whether pursuing goals concerning the output variables contributes to the semantic effect we devised two additional versions of the same system that differ with regard to their proneness towards goal adoption whilst keeping semanticity constant. In the *MACHINE-output* version of the problem the input variables were labelled "Light", "Water" and "Temperature" whilst the output variables carried *MACHINE* labels, i.e. "X", "Y" and "Z". In the *CHERRY-TREE-output* version the input variables were labelled "A", "B" and "C", whilst the output variables were labelled "Cherries", "Leaves" and "Beetles". Underpinning were two assumptions: (1) goal adoption is focused on output variables, and (2) goal adoption is most likely to occur for output variables high in semanticity. In other words, pursuing to "increase number of CHERRIES, keep number of LEAVES constant, and decrease number of BEETLES", as an example, is more meaningful than aspiring to "increase X, keep Y constant, and decrease Z". Hence we hypothesise that if goal adoption was the primary reason for the semantic effect then its adverse effects on knowledge acquisition should be observable in conditions with *CHERRY TREE*-related labels for output variables and less so in conditions with *MACHINE*-related labels for output variables (*Goal adoption hypothesis: knowledge acquisition*). From a presumption perspective no such differences would be expected between these two conditions as there is no reason to assume that the

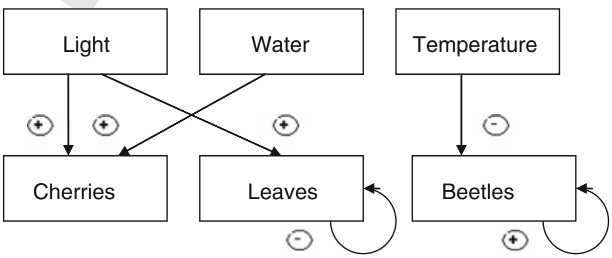

Fig. 1 Diagram of the causal structure of the dynamic system

Journal : Small 11251 Dispatch : 31-5-2013 Pages : 20
Article No. : 9280 LE TYPESET
MS Code : CP DISK

combination of input variables high in semanticity and output variables low in semanticity would provoke more a priori assumptions than the reverse combination with input variables low in semanticity and output variables high in semanticity. In other words, we experimentally manipulated the likelihood of goal adoption by systematically varying the control worthiness of the output variable labels. The success in controlling such complex, dynamic system (which in this context means to reach and maintain pre-determined target values in the output variables over a certain period of time) depends on the quality of the knowledge acquired during the goal free exploration phase (Goode and Beckmann 2010). We, therefore, hypothesise that differences in control performances will be observable between these two conditions if goal adoption is causal to the semantic effect (*Goal adoption hypothesis: system control*). Again, from a presumption perspective we would not expect systematic control performance differences between these two conditions.

Tests of the presumption explanation for the semantic effect are based on the assumption that the semanticity of variable labels has an impact on problem solvers' a priori mental models of the system. We therefore expect problem solvers to list higher numbers of a priori assumptions regarding the interdependency of system variables if labelled with a *CHERRY TREE* context (*Presumption hypothesis:* a priori *assumptions*). If presumptions were to be responsible for the semantic effect we should observe performance differences in the acquisition of knowledge about the causal structure of the system between individuals who report high or low numbers of a priori assumptions, respectively (*Presumption hypothesis: knowledge acquisition*). This, consequently, should also be reflected in performance differences in their system control performance (*Presumption hypothesis: system control*).

Table 1 provides an overview of the allocations of the four system versions to the experimental conditions regarding control worthiness and semanticity.

Measures

Our analyses were based on four measures: the number of a priori assumptions, accuracy of knowledge acquired, systematicity of exploration interventions and quality of system control.

A priori assumptions

To assess participants' initial mental model of the structure of the system, a priori assumptions with regard to the existence (and non-existence) of each of the 12 possible relationships between the system variables were recorded. For that purpose, a template with drop down boxes was presented on screen (see bottom section in Fig. 2). The template elicited information regarding problem solvers' assumptions regarding the existence or non-existence of relationships between any given input variable and any given output variable, or a dependency of any of the three output variables on itself by selecting "Y" (yes) or "N" (no). For assumed relationships problem solvers then could also specify their

Table 1 Experimental factors across conditions

Experimental factors	*CHERRY TREE*	*CHERRY TREE out*	*MACHINE out*	*MACHINE*
Semanticity	High	Medium	Medium	Low
Control worthiness	High	High	Low	Low

Journal : Small 11251 Dispatch : 31-5-2013 Pages : 20
Article No. : 9280 ☐ LE ☐ TYPESET
MS Code : ☑ CP ☑ DISK

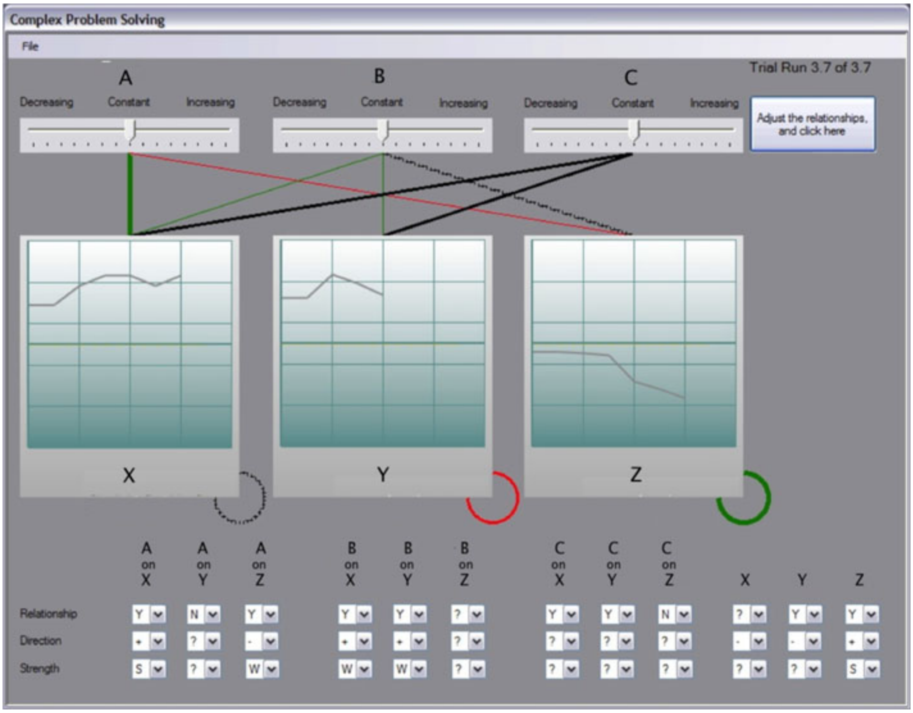

Fig. 2 Screen-shot of the task, as presented in the *MACHINE* condition

direction, by selecting "+" (positive) and "–" (negative). Finally, assumptions regarding the strength of the relationship were recorded, based on the selection of "W" (weak), "M" (medium) or "S" (strong) in the drop down menus. This information was then used to generate a causal diagram, which was presented onscreen (see middle section in Table 2). Solid arrows represent links assumed to exist; dashed arrows represent links were no decision was made; absent arrows represent assumptions regarding non-existences of links. The a priori assumption score has a theoretical range of 0–12.

Accuracy of acquired knowledge

Participants' knowledge was assessed by asking them to complete causal diagrams of the structure of the problem before they began the control task, and after each exploration trial (see above mentioned procedure). The diagrams generated before they began the task and at the end of the three exploration cycles with seven trials each were used to derive problem solvers' knowledge scores. These scores represent sensitivity scores (*Pr*) adopted from memory recognition research (Snodgrass and Corwin 1988). In contrast to traditional signal detection models Snodgrass and Corwin's (1988) model conceptualises discrete states of recognition (rather than a continuum of memory strengths). Translated into the context of the acquisition of structural knowledge we distinguish between the states of either knowing (that there is or is not a relationship between two variables) or not knowing. This model also allows controlling for guessing (i.e. the tendency "to see that there is" or "to see that there is not" a relationship in the state of not knowing). *Pr* scores have a

Journal : Small 11251 Dispatch : 31-5-2013 Pages : 20
Article No. : 9280 ☐ LE ☐ TYPESET
MS Code : ☑ CP ☑ DISK

Table 2 Procedure of the experiment and performance measures collected

Phases	Variables
1. Instruction	—
2. Assessment of expectations	• Number of assumptions • Accuracy of prior knowledge (Pr_0)
3. Goal-free exploration (3 cycles with 7 trials each)	• Accuracy of knowledge acquired (Pr_1–Pr_{21}) • Systematicity (*VONAT*)
4. System control (1 cycle with 7 trials)	• Quality of control interventions (*Control*)

theoretical range from -1 to 1, where scores below zero indicate inaccurate knowledge, whilst scores above zero indicate accurate knowledge.

Systematicity of exploration

To determine whether participants explored the system systematically, the number of trials in which one or none of the three inputs were varied was recorded; the VONAT (i.e., vary one or none at a time) scores theoretically range from 0 to 21.[3]

System control

The quality of system control was calculated by determining the Euclidean Distance between the actual and optimal values of the input variables on each of the seven control trials during the control cycle. The optimal values for each input variable were calculated by using the values of the output variables on the previous trial and the target output values to solve the set of linear equations underlying the system. These scores are then averaged across the seven control trials. The theoretical range of the scores was from 0 to 121, where a lower score indicated a smaller deviation from optimal control interventions and therefore better performance.

Procedure

Table 2 shows the procedure of the experiment, and indicates which performance data were collected in each phase of the experiment. At the start of the experiment participants were verbally instructed how to interact with the system, aided by a PowerPoint presentation. The system was then presented on a computer. The entire experiment took under an hour to complete.

Before engaging in any interaction with the system, participants were required to enter any assumptions they had regarding the interconnectedness of the system variables. This information was then used to generate a causal diagram, which remained on screen (see middle section in Fig. 2).

The exploration phase then began, in which participants were prompted to learn about the causal structure of the system. To this end they were given 3 cycles of 7 trials each (i.e. a total of 21 trials) where they could change the three input variables (see sliders in the top

[3] Technically, only four interventions are necessary to completely identify a linear 3 by 3 system: one where none of the input variables are changed to identify autonomic changes in the output variables, and three interventions where only one of the input variables is changed in order to identify their respective effects on the output variables.

Journal : Small 11251 Dispatch : 31-5-2013 Pages : 20
Article No. : 9280 ☐ LE ☐ TYPESET
MS Code : ☑ CP ☑ DISK

section in Fig. 2) and then observe the effect on the output variables (see graph windows in the middle section in Fig. 2). After each trial, participants were required to record what they had learnt about the dependencies of the system variables using the template on screen. These inputs modified the causal diagram accordingly. The causal diagram serves as an externalisation of the acquired and to be refined mental model about the causal structure of the system. The feature of enabling problem solvers to update the causal diagram over the course of the 21 exploration trials accommodates the accumulative nature of knowledge acquisition.

In the control phase, which consisted of seven intervention trials, participants were instructed to control the system to reach and maintain given target values for the three output variables. The target values were marked by yellow lines in the graph windows for the output variables.

Analyses

A between-subjects design was used with a total of four system versions that allowed contrasting (a) three levels of semanticity and (b) two levels of control worthiness. The three levels of semanticity (i.e. low, medium, and high) aimed at the experimental induction of varying levels of proneness towards the development of a priori assumptions. The low level of semanticity is represented by the system version that comprises exclusively of *MACHINE* related variable labels, the two system versions that had either input or output variables with *CHERRY TREE* labels represented the medium level of semanticity. The system version with *CHERRY TREE* labels for both, input and output variables represented high levels of semanticity (cf. Table 1). To test the presumption hypotheses we first established whether a link existed between different levels of semanticity of a system and the amount of a priori assumptions problem solvers held. Subsequently, we contrasted performance measures obtained under varying levels of presumptiveness using ANOVA.

Testing the goal adoption hypotheses required an experimental manipulation of the control worthiness of a system resulting in different levels of proneness towards goal adoption. This was achieved by contrasting the two system variants with output variable labels high in control worthiness (i.e. *CHERRY TREE* context) with the two system variants where output variable labels are not likely to encourage goal adoption (i.e. *MACHINE* related output variables). The effects of control worthiness on knowledge acquisition as well as control performance were analysed using ANOVA.

Results

The random allocation of participants to the four experimental conditions resulted in the following groups: $n_{MACHINE} = 21$ (14 females), $n_{CHERRY\ TREE} = 20$ (15 females), $n_{CHERRY\ TREE\text{-}output} = 20$ (15 females) and $n_{MACHINE\text{-}output} = 19$ (14 females).

Table 3 provides descriptive statistics of the variables considered in the analyses. The last column in Table 3 shows data derived from $n = 20$ generated data sets with random responses to the a priori assumption assessment and to the knowledge assessment as well as random control interventions.

The reliability estimation for the performance measures (i.e. knowledge acquisition scores and control performance scores) were based on the data from all 80 participants, regardless of the condition under which they dealt with the complex, dynamic system. Cronbach's alpha for the first, second, and third set of seven trials (i.e. exploration cycles)

Journal : Small 11251 | Dispatch : 31-5-2013 | Pages : 20
Article No. : 9280 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

Table 3 Descriptive statistics for the dependent variables [$M(SD)$]

Variables	*CHERRY TREE (n = 20)*	*CHERRY TREE out (n = 20)*	*MACHINE out (n = 19)*	*MACHINE (n = 21)*	*Random (n = 20)*
A priori assumptions	5.75 (5.39)	2.85 (4.89)	4.74 (5.51)	2.05 (4.12)	7.25 (1.80)
Accuracy of knowledge based on assumptions: Pr_0	−.008 (0.15)	.025 (0.10)	−.052 (0.13)	−.016 (0.10)	−.008 (0.28)
Systematicity of exploration interventions: *VONAT*	9.55 (6.31)	9.05 (7.35)	9.42 (7.21)	12.43 (4.75)	–
Accuracy of acquired knowledge: Pr_{21}	.041 (0.18)	.230 (0.27)	.173 (0.31)	.328 (0.32)	.008 (0.26)
Control performance: *Control*	40.05 (12.11)	39.10 (18.26)	41.08 (11.19)	31.07 (13.26)	53.26 (12.06)
$r(Pr_{21}, Control)$	−.16	−.14	−.17	−.52	.14

were .81, .97 and .98, respectively. Cronbach's alpha for all 21 interventions during the exploration phase was .97. The internal consistency for the seven interventions in the control phase was also sufficiently high, resulting in a Cronbach's alpha of .89.

Semantic effect

Prior to testing the specific predictions that were put forward in the contexts of the goal adoption or presumption hypotheses, we tested whether the "semantic effect" was replicated in the current study. To this end, we compared the knowledge scores (i.e. Pr_{21}, the final score after 21 exploration trials) achieved in the *CHERRY TREE* and in the *MACHINE* version. To control for potential differences in their "a priori knowledge" (i.e. the knowledge score based on the a priori assumptions elicited from each problem solver prior to their first exploration trial), Pr_0 was included as a covariate. The result confirmed the expected replication of the semantic effect ($F_{1,38} = 12.94$, $p = .001$, $\eta^2 = .248$). Problem solvers in the different conditions started off at comparable levels of "a priori knowledge" (effect of covariate Pr_0: $F_{1,3} = 1.52$, $p = .225$). However, the semanticity of the system variables had an effect on how much problem solvers learnt about the structure of the system. Problem solvers dealing with labels with high semanticity tended to acquire less knowledge than problem solvers working under the low semanticity condition. In fact, when compared with knowledge scores derived from 20 simulated data sets with random causal diagrams (representing guessing or "zero knowledge") problem solvers' knowledge in the high semanticity condition (i.e. *CHERRY TREE*) did not differ from that ($F_{1,37} = 0.24$, $p = .631$).

In a subsequent step we had to establish whether poor knowledge acquisition under high semanticity conditions is attributable to an incompatibility between "common sense" (as indicated by the a priori assumptions about the structure of the system held by problem solvers) and the actual structure of the system (see Fig. 1 for the underlying causal structure). If this were the case, the Pr_0 scores in the *CHERRY TREE* condition should be significantly less than zero, signalling systematic "false knowledge", which would be indicative of a counterintuitive system structure. The results of a one-sample t test demonstrated however that Pr_0 scores in this condition ($M = -0.008$, $SD = 0.15$) were not significantly different from zero ($t_{[19]} = -0.252$, $p = .804$) which indicates that problem solvers, on average, held an equal number of correct and incorrect assumptions with regard to the actual structure of the system. This eliminates the possibility that the "semantic effect" might have simply been caused by a counterfactual causal structure.

As to be expected, the semantic effect was replicated also with regard to the control performance. The significant effect of semanticity ($F_{1,39} = 5.11$, $p = .029$, $\eta^2 = .116$) in an ANOVA contrasting the control performance between the *CHERRY TREE* and the *MACHINE* version of the system signifies that problem solvers' control performance under conditions with low semanticity (i.e. *MACHINE*) was superior to those working under high semanticity conditions (i.e. *CHERRY TREE*). However, although little knowledge was acquired during the exploration phase, the control performances under *CHERRY TREE* conditions did significantly differ from scores derived from 20 simulated data sets with random control interventions, representing some sort of "knowledge free control" ($F_{1,38} = 11.95$, $p = .001$, $\eta^2 = .239$).

Goal adoption hypotheses

To test whether the premature pursuance of controlling the system (i.e. goal adoption) can explain the semantic effect we contrasted knowledge acquisition performance as well as control performances achieved when working with system versions that vary with regard to the control worthiness of output variables. We compared performance scores from the system versions with *CHERRY TREE* output variable labels (i.e. high in control worthiness) with the system versions with *MACHINE* output labels (i.e. low in control worthiness). The results indicate that there is no significant difference in the final knowledge acquisition score (Pr_{21}: $F_{1,78} = 3.48$, $p = .066$[4]) between these two groups. This result pattern was replicated for the comparison of control performance (control: $F_{1,78} = 1.38$, $p = .244$). These results do not support either of the two Goal adoption hypotheses (i.e. knowledge acquisition and system control).

Presumption hypotheses

To test the alternative explanation for the semantic effect, we first analysed the distribution of the number of a priori assumptions problem solvers held across the three different levels of semanticity. In general, we expected to find an increase in the number of a priori assumptions with increasing levels of semanticity. As shown in Fig. 3 this expectation was confirmed, which lends tentative support to the presumption explanation (Presumption hypothesis: a priori assumptions).

An inspection of the distribution of the number of presumption across the levels of semanticity also reveals bi-modality. This suggests that problem solvers tend either to have none or only a few presumptions (i.e. less than 4) or they have many assumptions (i.e. eight or more). In light of these findings we created a quasi-experimental factor with two levels of a priori assumptions. All problem solvers with less than 6 (out of possible 12) a priori assumptions were allocated to the "low assumption" group ($n_{low} = 53$); all problem solvers with six or more assumptions formed the "high assumption" group ($n_{high} = 27$).

The calculation of Somers' *D* (Somers 1962), as an estimate of the size of the assumed directional association between an ordinal and a dichotomous variable, resulted in a value of .246 ($p = .007$) indicating a significant dependency of the number of a priori

[4] Under given sample size constellations an existing effect of at least medium size (i.e. $d \geq 0.57$)—which in this form of analysis also translates into an effect of as small as 9 % of explained variance—will be detectable with a probability of more than .80.

Journal : Small 11251 | Dispatch : 31-5-2013 | Pages : 20
Article No. : 9280 | LE | TYPESET
MS Code : | CP | DISK

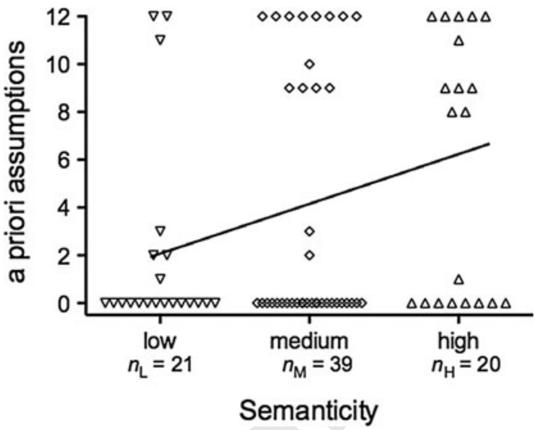

Fig. 3 Number of a priori assumptions across different levels of semanticity

assumptions from the level of semanticity. Somers' *D* can be interpreted as the difference between two probabilities (Newson 2006, p. 311), which in the given context means that the probability to adopt high levels of assumptions will increase by .25 when confronted with higher levels of semanticity (see *Presumption hypothesis:* a priori *assumptions*).

Presumption hypothesis: knowledge acquisition

An ANCOVA analysing the effect of a priori assumptions on knowledge acquisition (Pr_{21}), whilst controlling for potential differences in Pr_0, resulted in a significant main effect ($F_{1,77} = 6.13$, $p = .016$, $\eta^2 = .073$), indicating that problem solvers with a high number of a priori assumptions tend to acquire less knowledge over the course of the exploration phase. Regardless of whether problem solvers held high or low numbers of assumptions, they did not differ in their "a priori knowledge" as signified by the non-significant effect of the covariate Pr_0 ($F_{1,77} = 0.003$, $p = .958$).

Presumption hypothesis: system control

As expected, the subsequent control performances also differed significantly between problem solvers with high and low numbers of a priori assumptions ($F_{1,78} = 24.60$, $p < .001$, $\eta^2 = .240$).

The accumulated evidence regarding systematic differences in knowledge acquisition and control performance between the two assumption groups lends further support to the presumption explanation. However, holding a priori assumptions about the system structure is not necessarily a hindrance to knowledge acquisition and the subsequent reason for inferior control performance per se. Only if not systematically tested during the exploration phase assumptions impose a consequential threat to the acquisition of new knowledge. An ANOVA with the two assumption levels (high vs. low) and the individual degree of systematicity of interventions across the three exploration cycles (i.e. *VONAT*) was conducted. The significant effect of assumptiveness ($F_{1,78} = 30.62$, $p < .001$, $\eta^2 = .282$) confirms that high levels of assumptions are substantially associated with low levels of systematicity in exploration behaviour. As to be expected, systematicity in exploration behaviour, on the other hand, is positively associated with accuracy of acquired knowledge ($r = .32$, $p = .002$).

Overall, the results with regard to the presumption hypotheses indicate that a problem context with high levels of semanticity tend to encourage the formation of higher numbers of a priori assumptions. These assumptions are less likely to be tested systematically as indicated by the significantly lower level of systematcity in the exploration behaviour of this group. Unsystematic exploration behaviour tends to produce non-informative system states (i.e. design of "inconclusive experiments", de Jong and van Joolingen 1998, p. 185) that complicate the extraction of knowledge (i.e. induction of causal structure). A lack of knowledge ultimately leads to poor control competency.

Discussion

The goal of this study was to better understand how the semantic context of a problem might influence the way problem solvers approach a complex, dynamic system. We were interested in further elucidating the underlying mechanisms that lead to the semantic effect. To that end we explored the viability of two—potentially alternative—explanatory mechanisms, namely goal adoption and presumptions.

The study replicates results reported by Beckmann (1994; Beckmann and Guthke 1995) with regard to the semantic effect, which refers to the phenomenon of impeded acquisition of knowledge about the causal structure of a complex, dynamic system as well as inferior control performance when presented with semantically familiar labels in contrast to the use of abstract variable labels.

In terms of the quest for an explanatory mechanism for this phenomenon, empirical evidence obtained in this study lends little support for goal adoption as the underlying cause. Contrary to what would be expected if goal adoption were causally linked to the semantic effect, we were not able to find systematic performance differences (i.e. knowledge acquisition and system control) between conditions presumably most and least conducive to goal adoption. This could mean two things: either goal adoption does not result in a semantic effect or goal adoption did not occur. We do not have a direct indicator of whether problem solvers in conditions with high levels of control worthiness have indeed engaged in pursuing self-set goals. Systematicity of interventions could serve as an, although indirect indicator. In this sense, we would argue that inputs that are aimed at exploring the causal structure of a system are more systematic in comparison to interventions targeted at reaching and maintaining certain target values in output variables. Systematic exploration behaviour focuses on individual links between a single input and a single output variable. To establish the existence of such links interventions are required where only one input variable is changed whilst the other input variables are be kept constant (see VONAT). Control goals, in contrast, refer to all output variables simultaneously. Therefore, control interventions require changes of more than just one input variable. As a result, input behaviour typical of goal adoption should be characterised by lower VONAT scores. A comparison of VONAT scores for *CHERRY-TREE-output* and *MACHINE-output* conditions (see Table 3) suggests no such difference. If we accept systematic differences in VONAT scores as an (indirect) indicator of goal adoption and if we accept that goal adoption is less likely occur in a context of variables low in semantictity we have to conclude that goal adoption might not have taken place in our study.

Tests of the presumption explanation resulted in a more coherent result pattern. The likelihood of adopting higher numbers of a priori assumptions increases by 25 % when being confronted with a system with high levels of semanticity. Problem solvers tend not to systematically test these presumptions, i.e. 28 % of variation in the systematicity of

Journal : Small 11251 | Dispatch : 31-5-2013 | Pages : 20
Article No. : 9280 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

decisions made during the goal free exploration phase can be accounted for by the level of a priori assumptions. Higher levels of systematicity led to more accurate knowledge about the causal structure of the system, which ultimately enabled problem solvers to better control the complex, dynamic system. Overall, the level of presumptions explains 7 % in the variation of knowledge acquisition scores and 24 % in control performance, respectively. All in all, it is the combination of both, semanticity induced presumptions and failing to test them systematically that qualifies as a *sufficient cause* for the semantic effect. Consequently, the semantic effect should be more appropriately referred to as "presumption effect". As long as we cannot demonstrate an occurrence of the semantic effect independently of untested presumptions, they also have to be considered a *necessary cause*.

On the other hand, if we accept that goal adoption might not have occurred in this study then goal adoption has to be ruled out as a necessary cause for the semantic effect. For the same reason, we cannot decide whether goal adoption could be a sufficient cause for the semantic effect. The fact that goal adoption might not have occurred in this study—despite creating arguably conducive conditions for it—raises doubts regarding its viability as a "naturally" occurring phenomenon. Goal adoption might only occur if problem solvers were informed about goals before they are asked to explore the system.

Before discussing some of the implications of these findings with regard to complex problem solving research and to appropriately map the findings into the field of instructional design we briefly bring to mind the specifics of the learning task imposed in this study. Strictly speaking, the task was twofold. The first subtask was a learning task in a narrow sense, whilst the second subtask can be seen as an indirect learning task. The first subtask required the acquisition of knowledge about the causal structure of a complex, dynamic system. The second subtask comprised of the application of the acquired knowledge to reach a defined target. In the context of cognitive learning, the knowledge acquisition task required the induction of rules regarding the interdependencies of variables in this system from a series of observations. These observations had to be generated by the problem solver through systematic interactions with the system. Systematicity in these interactions pretty much is aligned with systematic inquiry used in experimentation (e.g. change one variable at a time whilst keeping others constant) to draw causal inferences. The second subtask builds on the first in so far as the mental model developed during the preceding exploration phase (see subtask 1) has to be utilised to control the system. System control, in this context, means reaching and maintaining a goal state of the system that comprised certain target values for a subset of system variables. Performance in both subtasks, i.e. knowledge acquisition and system control, depends on a range of intrapersonal and situational factors. Reasoning ability would be an example for the former, semantic context of the problem would be an example for the latter. This study was concerned with the impact of semanticity on learning and the application of acquired knowledge.

We now discuss some of the implications of these findings with reference to two broad contexts, one is the use of complex, dynamic problems in research on cognitive learning, and the other one refers to instructional design.

The use of complex dynamic systems in research in cognitive learning, reasoning and intelligence

Computer-simulated scenarios are prominent tools used in the study of human problem solving behaviour. As it was in this study, the focus is on how people learn to operate in complex, dynamic environments. One of our findings suggests—in accordance with previous studies—that a complex, dynamic problem can be controlled at levels substantially better than pure random interventions without the acquisition of (explicable) knowledge about its causal structure. In the context of the apparent dissociation between control competency and knowledge the imperfection of the methods used to assess the results of knowledge acquisition processes led to the notion of implicit, i.e. non-explicable knowledge or implicit learning. As in Goode and Beckmann (2010) we argue that the role of implicit learning is negligible in the given context. More likely are processes of what has been described previously as "ad hoc control" (Beckmann and Guthke 1995, p. 195). "Ad hoc control" refers to an intervention-by-intervention optimization, or simply trial-and-error-processes during system control. As an example, if the first ("knowledge free") intervention caused the system to deviate further from the goal states then a reasonable ad hoc control strategy to (over-)compensate this undesired development would be to enter values for the input variables that carry the algebraic sign opposite to the previous intervention trial. Inputs that tend to bring the system closer to the goal state are likely to be repeated.

The heuristic of "ad hoc control" does not require any structural knowledge about the system. The correlation pattern between knowledge acquisition and control performance (see bottom row in Table 1) in combination with the knowledge acquisition and control performance scores across the different conditions supports this claim.

One of the implications is that sole reliance on control performance as an indicator of learning success is inadequate. The potential multiple determination (i.e. ad hoc, knowledge based etc.) of control performance imposes a validity challenge to the use of control performance in complex, dynamic problems as a proxy for knowledge and learning.

We see another consideration that speaks against the reliance on control performance measures in the context of research on learning, especially learning processes. This concern is most relevant for task settings where problem solvers are given control goals without a preceding goal free exploration or learning phase, exclusively dedicated to the acquisition of structural knowledge. Situations in which we expect problem solvers to learn whilst they are in fact being asked to perform (e.g. to control a system)—occasionally disguised as "learning by doing"—are less likely to promote effective learning as a systematic testing of assumptions are seldom expedient to reaching control targets. Again, when studying learning behaviour in complex problem scenarios we therefore should not rely on performance measures such as control performance to make inferences regarding the quality of knowledge acquisition.

The presumption effect in relation to instructional design

In our pedagogical attempts to make problems or learning tasks more interesting, more relevant, more meaningful and subsequently more motivating to problem solvers, we should not forget that the references to real life systems as suggested by variable labels are of a mere token nature. As shown in this study, problem solvers have difficulties in treating certain kind of information as an irrelevance. Although in a different context, Cooper and colleagues gave a critical account of the risk associated with the use of potentially confusing "realistic" items in their research on knowledge and skill acquisition in mathematics (Cooper and Dunne 2000; see also Lubienski 2000). Variable labels are quite tempting as they offer a (too) convenient way of sense making that is hard to resist. Goldstone and Son (2005, p. 100) refer to "people's natural tendency to interpret ambiguous objects so as to be consistent with (previously experienced), unambiguous objects …" In the context of learning tasks imposed in the present study we would argue that the use of (overly) concrete labels (i.e. high semanticity) creates a situation where problem solvers are likely to be tempted to fill the "ambiguity vacuum" with assumptions that are not necessarily incorrect (see neutral knowledge scores based on a priori assumptions) but not necessarily helpful either. In short, the use of systems with high semanticity comes with the risk of diverting problem solvers from a learning path.

In our study, we observed problem solvers who resisted the temptation to adopt high levels of a priori assumptions under conditions with high semanticity (i.e. 9 out of the 20, or 45 %, see Fig. 3). At the same time, the adoption of high levels of assumptions in the low semanticity condition is considerably lower (i.e. 3 out of 21, or 14 %, see Fig. 3). Not adopting high levels of assumptions can be seen as the first protective factor against the semantic effect. As said before, high levels of a priori assumptions are not detrimental to knowledge acquisition per se; failing to test them systematically is.

The consistently observed difficulties problem solvers seem to experience in systematically testing their a priori assumptions poses a challenge to constructivist, discovery, problem-based, experiential, and inquiry-based teaching (Kirschner et al. 2006; Klahr and Nigam 2004). Our results suggest that it is somewhat presumptuous to rely on hypothesis testing in problem solving to take place "naturally" (Gopnik et al. 2001). As a recommendation implied by our results, rather than condemning the use of systems with high levels of semanticity, it appears crucial to provide problem solvers with guidance on how (a) to explicate assumptions and (b) to test them systematically.

Research in the context of discovery learning using simulations (e.g. de Jong and van Joolingen 1998) has identified a range of similar difficulties learners tend to experience which include *hypothesis generation*, *design of experiments*, *interpretation of data*, and *regulation of learning*. Subsequently, a variety of remedial support features or pedagogies are proposed. They, for example, include a program feature called "hypothesis scratchpad" (van Joolingen and de Jong 1993) that is meant to encourage the generation of hypotheses. Also, more or less explicit hints as to what kind of interaction with the computerised system (i.e. inputs) are likely to generate most informative system states were suggested (Leutner 1993). Another attempt to support learners is to ask for explicit predictions of outcomes of certain interventions (Beckmann 1994), which, in their comparison to the actual outcomes, is expected to generate feedback informative of whether the assumed (hypothesised?) mental model is correct or needs modification. These support features, however, have not consistently shown the expected effects (de Jong and van Joolingen 1998, p. 193).

Despite their differences we see relevance of our findings with regard to discovery learning at a conceptual as well as methodological level. Discovery learning is mainly concerned with concept learning (employing *conceptual models* in simulations). The study presented here focuses on the application of a generic problem solving skill (i.e. drawing causal inferences based on systematic experimentation) which utilises a computerised system that follows an *operational model*). On a conceptual level, the findings reported here suggest that the effectiveness of discovery learning is potentially threatened by one of its core features: its reliance on contextual embedment. On a methodological level it is apparent that in the majority of relevant studies in discovery learning differences in exploration behaviour between successful and less successful learners are used as the basis for predominantly descriptive accounts of difficulties learners tend to experience. The study presented here, however, goes beyond the ex post facto comparison of successful and less successful learners by focussing on situational characteristics (i.e. semantic embedment as a design feature of contextualised learning) that may contribute to learners' difficulties in discovery learning using computerised scenarios.

A further potential problem of using highly contextualised learning environments is that learning products (i.e. knowledge and understanding) are tied to specific contexts. Knowledge and skills acquired within narrowly defined contexts are less likely to be transferred into novel, yet homomorphous situations. To mitigate this problem Goldstone and Son (2005) convincingly proposed the approach of *concreteness fading* (see also Schwartz and Black 1996) which refers to "… the process of successively decreasing the concreteness of a simulation with the intent of eventually attaining a relatively idealized and decontextualized representation that is still clearly connected to the physical situation that it models" (Goldstone and Son 2005. p. 70). Learning and transfer are best facilitated by moving from higher to lower levels of semanticity and not by the reverse sequence as Goldstone and Son (2005) as well as Koedinger and Anderson (1998) were able to show.

This raises the question whether our findings stand in contradiction to the concreteness fading approach. Not necessarily. A possible reconciliation lies in differences in the actual learning tasks. The learning task set in Goldstone's studies is aimed at developing an understanding of a specific scientific principle or concept (i.e. "competitive specialisation"—a principle that underpins the decentralised organisation of complex behaviour such as bird flocks, traffic jams and forest fires, see Resnick 1994). As stated before, the learning task posed in the study we have presented here aimed at a rather generic problem solving skill that is central to scientific enquiry, namely drawing causal inferences based on systematic experimentation. Successful experimentation represents the basis for an understanding of principles such as "competitive specialisation". Hence, we argue that the former functionally serves as a precursor for the latter. A synthesis of the two seemingly contradictory pieces of evidence results in the insight that different learning foci call for different features in learning environments (such as semantic embedment) as do different phases in the process of acquiring knowledge.

Desiderata

The next questions to address are: (a) Why are presumptions not tested systematically by default; and (b) What can be done about it. Tentative answers to these questions are that semanticity tends to encourage a (false) sense of familiarity or an "illusion of knowing". Furthermore, starting a learning process with a set of assumptions ideally requires a systematic *reduction of complexity* of the initial mental model of the problem. Problem contexts with low levels of semanticity are more likely to make it easier for the problem solver to realise that they might not know. Hence the necessity to acquire knowledge, rather than relying on presumptions and leaving them untested, comes to the fore. Learning and knowledge acquisition under these circumstances represents a process of building up or *creation of complexity*.

When creating learning environments we need to ensure that cover stories and variable labels—as well-intended as they might be—are not standing in the way of learning. If ignored by using semantically embedded problems uncritically at least two parties will be negatively affected; the researcher who is interested in studying learning processes and the learner who is confronted with this kind of material.

Acknowledgments This research was supported, in part, under the Australian Research Council's Linkage Projects funding scheme (project LP0669552). The views expressed herein are those of the authors and are not necessarily those of the Australian Research Council.

Journal : Small 11251 | Dispatch : 31-5-2013 | Pages : 20
Article No. : 9280 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

References

Beckmann, J. F. (1994). *Learning and complex problem solving*. Bonn: Holos.

Beckmann, J. F., & Guthke, J. (1995). Complex problem solving, intelligence, and learning ability. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 177–200). Hillsdale, NJ: Erlbaum.

Blessing, S. B., & Ross, B. H. (1996). Content effects in problem categorisation and problem solving. *Journal of Experimental Psychology, 22*, 792–810.

Burns, B. D., & Vollmeyer, R. (2002). Goal specificity effects on hypothesis testing in problem solving. *The Quarterly Journal of Experimental Psychology, 55A*, 241–261.

Chung, G., Harmon, T. C., & Baker, E. L. (2001). The impact of a simulation-based learning design project on student learning. *IEEE Transactions on Education, 44*, 390–398.

Cooper, B., & Dunne, M. (2000). *Assessing children's mathematical knowledge: Social class, sex and problem-solving*. Buckingham: Open University Press.

de Freitas, S., & Oliver, M. (2006). How can exploratory learning with games and simulations within the curriculum be most effectively evaluated? *Computers & Education, 46*(3), 249–264.

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

Dörner, D. (1987). On the difficulties people have in dealing with complexity. In J. Rasmussen, K. Duncan, & J. Leplat (Eds.), *New technology and human error*. New York: Wiley.

Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science, 17*(3), 397–434.

Fang, L., Tan, H. S., Thwin, M. M., Tan, K. C., & Koh, C. (2011). The value simulation-based learning added to machining technology in Singapore. *Educational Media International, 48*, 127–137.

Funke, J. (1992). Dealing with dynamic systems: Research strategy, diagnostic approach and experimental results. *German Journal of Psychology, 16*(1), 24–43.

Goldstone, R. L., & Sakamoto, Y. (2003). The transfer of abstract principles governing complex adaptive systems. *Cognitive Psychology, 46*, 414–466.

Goldstone, R. L., & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. *The Journal of the Learning Sciences, 14*, 69–110.

Goode, N., & Beckmann, J. F. (2010). You need to know: There is a causal relationship between structural knowledge and control performance in complex problem solving tasks. *Intelligence, 38*, 345–352.

Gopnik, A., Sobel, D., Schulz, L., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two, three, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37*(5), 620–629.

Greiff, S., & Funke, J. (2009). Measuring complex problem solving—The MicroDYN approach. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: Lessons learned from large-scale surveys and implications for testing* (pp. 157–163). Luxembourg: Office for Official Publications of the European Communities.

Guthke, J., Beckmann, J. F., & Stein, H. (1995). Recent research evidence on the validity of learning tests. In J. S. Carlson (Ed.), *Advances in cognition and educational practice. European contributions to the dynamic assessment* (Vol. 3, pp. 117–143). Greenwich: JAI Press.

Hesse, F. W. (1982). Effekte des semantischen Kontexts auf die Bearbeitung komplexer Probleme [Effects of semantic context on complex problem solving]. *Zeitschrift für Experimentelle und Angewandte Psychologie., 29*(1), 62–91.

Hesse, F. W., Kauer, G., & Spies, K. (1997). Effects of emotion-related surface similarity in analogical problem solving. *American Journal of Psychology, 110*, 357–385.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75–81.

Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes*. Cambridge, MA: MIT Press.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science, 15*(10), 661–667.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*(2), 211–228.

Koedinger, K. R., & Anderson, J. R. (1998). Illustrating principled design: The early evolution of a cognitive tutor for algebra symbolization. *Interactive Learning Environments, 5*, 161–180.

Kotovsky, K., & Fallside, D. F. (1989). Representation and transfer in problem solving. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 69–108). Hillsdale, NJ: Erlbaum.

Kröner, S., Plass, J. L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence, 33*, 347–368.

Lainema, T., & Nurmi, S. (2006). Applying an authentic, dynamic learning environment in real world business. *Computers & Education, 47*, 94–115.

Lazonder, A. W., Wilhelm, P., & Hagemans, M. G. (2008). The influence of domain knowledge on strategy use during simulation-based inquiry learning. *Learning and Instruction, 18*, 580–592.

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

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.

Lubienski, S. T. (2000). Problem solving as a means toward 'Mathematics for All': An exploratory look through a class lens. *Journal for Research in Mathematics Education, 31*, 454–482.

McGaghie, W. C., Issenberg, S. B., Petrusa, E. R., & Scales, R. (2006). Effect of practice on standardised learning outcomes in simulation-based medical education. *Medical Education, 40*, 792–797.

Newson, R. (2006). Confidence intervals for rank statistics: Somers' D and extensions. *The Stata Journal, 6*, 309–334.

Njoo, M., & de Jong, T. (1993). Exploratory learning with a computer simulation for control theory: Learning processes and instructional support. *Journal of Research in Science Teaching, 30*(8), 821–844.

Omodei, M. M., & Wearing, A. J. (1995). The fire chief microworld generating program: An illustration of computer-simulated microworlds as an experimental paradigm for studying complex decision-making behavior. *Behavior Research Methods, Instruments & Computers, 27*, 303–316.

Ravert, P. (2002). An integrative review of computer-based simulation in the education process. *Computers, Informatics, Nursing, 20*, 203–208.

Resnick, M. R. (1994). *Turtles, termites, and traffic jams*. Cambridge, MA: MIT Press.

Schauble, L., Klopfer, L. E., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching, 28*(9), 859–882.

Schoppek, W. (2002). Examples, rules, and strategies in the control of dynamic systems. *Cognitive Science Quarterly, 2*, 63–92.

Schwartz, D. L., & Black, J. B. (1996). Shuttling between depictive models and abstract rules: Induction and fallback. *Cognitive Science, 20*, 457–497.

Snodgrass, J., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology, 117*, 34–50.

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review, 27*, 799–811.

van Joolingen, W. R., & de Jong, T. (1993). Exploring a domain through a computer simulation: Traversing variable and relation space with the help of a hypothesis scratchpad. In D. Towne, T. de Jong, & H. Spada (Eds.), *Simulation-based experiential learning* (pp. 191–206). Berlin: Springer.

Vollmeyer, R., Burns, B. D., & Holyoak, K. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science, 20*, 75–100.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology*. Harmondsworth: Penguin.

Wood, R. E., Beckmann, J. F., & Birney, D. (2009). Simulations, learning and real world capabilities. *Education + Training, 51*(5/6), 491–510.