The German Journal of Psychology • Vol. 16 • No. 1 • 24–43 (1992)

Report

Dealing with Dynamic Systems: Research Strategy, Diagnostic Approach and Experimental Results[1]

Joachim Funke

Department of Psychology, Bonn University (FRG)

The method of computer-simulated scenarios has recently been introduced to study how people solve complex problems. This article describes a special approach to constructing such microworlds by means of linear structural equation systems. The subjects' task is to first identify in a knowledge acquisition phase the causal structure of a hitherto unknown system. In a later knowledge application phase they try to control this system with respect to a given goal state. Verbalizable knowledge that was acquired on the task is assessed both by means of causal diagrams as well as by the degree of successful control performance. Five experiments on special attributes of such systems illustrate the approach. The experiments investigated effects of active interventions versus observations only, effects of different degrees of Eigendynamik, the influence of different degrees of side effects, the role of prior knowledge, the amount of controllability and number of variables to be controlled. These factors have considerable effects on identification of the system structure and control of its states, these being two central indicators of complex problem solving. Three topics are identified as main goals for future research: (1) separation of different sources of variance (person, system, situation); (2) research on reliability and validity of performance indicators; (3) development of measures for an operators' heuristic and strategic knowledge.

1 Preparation of this article was supported in part by a grant from the "Deutsche Forschungsgemeinschaft (DFG)" to the author (Az. Fu 173/1). I wish to thank Gerd Fahnenbruck, Uwe Kleinemas, Horst Müller and Bärbel Rasche for their support in this research. Also thanks to Axel Buchner for improving the English text, and to Gerd Lüer and an anonymous reviewer who all gave valuable hints on an earlier version.

1. Introduction

Studying complex problem solving by means of computer-simulated scenarios ("microworlds") has become one of the favorite themes in German-speaking countries for researchers who are interested in the psychology of thinking. Instead of studying problem-solving behavior in restricted situations (such as "Tower of Hanoi" or "Cannibals and Missionaries" tasks; cf. Greeno, 1974; Jeffries, Polson & Razran, 1977), the new approach focuses on semantically rich domains that provide a touch of reality that was not inherent in previous research (see also Bhaskar & Simon, 1977; Broadbent & Aston, 1978; Mackinnon & Wearing, 1985). The main reason for shifting the emphasis to simulated real-life situations was to attain new insights into processes of thinking and action regulation under high cognitive and emotional load. Following the pioneering work of Dietrich Dörner (University of Bamberg, FRG) in the mid-seventies, several new scenarios have been developed and applied in correlational as well as in experimental studies (for a review see Funke, 1988). For instance, in a computer-administered microworld called "LOHHAUSEN," a subject has to take the role of an omnipotent mayor of a little town (Dörner, 1987). In other work, a subject plays the role of a manager of a little shop or of an engineer in a Third World country (Putz-Osterloh & Lemme, 1987). In general, the new approach deals with the exploration and control of complex and dynamic systems by human individuals. According to Dörner, subjects must satisfy the following requirements: (1) they must deal with the complexity of the situation and with the connectivity of the variables involved; (2) they must deal with the intransparency or opaqueness of the situation since typically not all information that is needed is available; (3) they must deal with dynamic developments of variables which change their states autonomously and make it necessary to anticipate trends; and (4) they must deal with multiple goals some of which may contradict others (e.g. as a manager: pay high wages due to the trade-union's request and at the same time maximize the company's profits).

Despite 10 years of research in the area, there is neither a clearly formulated specific theory nor is there any agreement on how to proceed with respect to the research philosophy. Even worse, no stable phenomena have been observed. Merely some rather general observations concerning some kind of breakdown of the cognitive system in cases of information overload have reliably been reported. Also, questions concerning the influence of test intelligence on complex problem solving have been answered ambiguously.

Therefore, the main concern of this article is to explain a systematic research strategy for studying how people deal with dynamic systems, and to illustrate the usefulness of this strategy with the help of experimental data. The paper will describe my own approach to studying complex problem solving which I conceive of as a dynamic process of knowledge acquisition and knowledge application in an information processing human. In Section

2, I will briefly describe the philosophy underlying this research, the DYNAMIS shell for creating and presenting scenarios, the general experimental procedure, the diagnostic approach to eliciting subjects' knowledge about the task, and the approach to measuring performance. In Section 3 the results of five studies within this framework will be presented. Finally, in Section 4 the results will be summarized and related to other studies. Also, I will give some perspectives for future research.

2. A Method for Analyzing Complex Problem Solving

From its beginning, research about solving complex problems had to cope with a number of difficulties (see the critical aspects mentioned by Funke, 1984). One central point was the reliable measurement of problem solving quality. Since no optimal solution path and no "best" intervention was available in most of these microworlds (because of the partially nonlinear relationships between the variables for which mathematically no optimal solution could be found), researchers were never quite sure whether a subject's solution to a problem was really better or worse than that of other subjects. However, at least qualitative judgments were possible. In other cases, in which subjects could set their own goals, problem solving quality was rated by "experts." Along with this came a complete loss of comparability of results. With these tasks it was impossible to separate out which part of the observed system changes was due to the tasks' characteristics and which was due to the subjects' attempt to cope with the problem. Also, the question of reliability of performance measures has been answered mainly by referring to the face validity of the tasks.

To overcome some of these problems, the line of research done in our Bonn laboratory established the following principles: (1) It should always be possible to define the quality of a solution by comparing it with an *optimal solution strategy*. (2) The situation should realize the *features of complex problems* (complexity, connectivity, intransparency, Eigendynamik [i.e., autonomous processes] and multiple goals) as far as possible. (3) A *detailed diagnostic procedure* should reveal subject's development of hypotheses about the system. This implies that subjects have to be prompted repeatedly about the causal structure they assume to the system. (4) There should be a clear distinction between a phase of *knowledge acquisition* (mainly realized by encouraging the subjects to explore the system) and a phase of *knowledge application* in which given states of the problem space should be reached by the subjects as quickly as possible.[2] In this last phase, performance measures should precisely indicate the quality of a subjects' intervention.

2.1 The DYNAMIS Shell for Scenarios

Trivially, before you can control a complex system, you must learn how it works. To study experimentally the acquisition as well as the application of knowledge we confront our subjects with computer-simulated scenarios. As a universal tool for constructing these scenarios a computer program called "DYNAMIS" serves as a shell with which the experimenter can implement in a simple way different types of simulated systems which all have in common one formal background. This general frame is a *linear equation system* (see e.g. Steyer, 1984) which consists of an arbitrary number of exogenous (= x) and endogenous (= y) variables according to the following equation:

$$y^{t+1} = A \times y^t + B \times x^t, \tag{1}$$

where y^{t+1} and y^t are vectors representing the state of the y-variables at times t+1 and t; x^t is a vector representing the values chosen by the subject for the x-variables; *A*, *B* are matrices containing the weights for the variables.

A set of measures for describing such systems formally has been suggested (e.g., Hübner, 1989). An equation system is constructed according to theoretical considerations about the presumed influence of certain system attributes on task complexity (e.g., the effect of Eigendynamik, or the influence of side effects or effects due to different interdependencies). It is not intended to simulate a domain of reality adequately because that kind of simulation places too many constraints on the attributes of the system to be useful for basic research on problem solving. Consequently, most of the simulated systems used in our research group have been "artificial." With respect to a distinction made by Hays and Singer (1989) one can say that what we want our systems to possess is not physical fidelity but rather functional fidelity. As an example see the SINUS system shown in Figure 1.

2 This distinction is somewhat artificial: In most cases knowledge acquisition occurs in order to reach certain goals. The goals in our settings concern the acquisition of the causal structure of the system under exploration with the goal of achieving control over that system (which is related to the application phase).

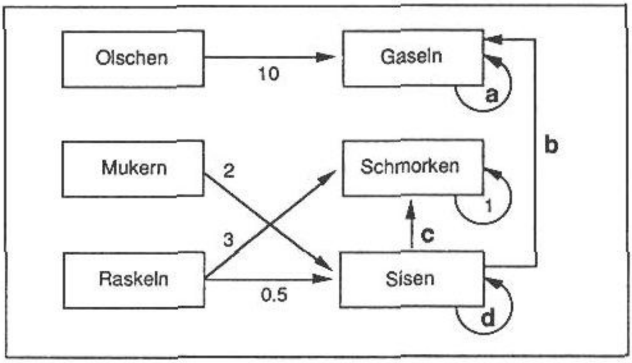

Figure 1. Causal structure of the system SINUS. The weight parameters a, b, c, and d are subject to experimental changes. The standard configuration is a = 1, b = 0, c = 0.2, and d = 0.9.

Subjects are told that this fictitious system consists of living creatures from a distant planet called SINUS. The "exogenous" variables are introduced as creatures labeled "Gaseln" (y_1), "Schmorken" (y_2) and "Sisen" (y_3), the "endogenous" creatures are called "Olschen" (x_1), "Mukern" (x_2) and "Raskeln" (x_3). The system has the following structure (parameters a, b, c, and d are variable weights):

$$y_1^{t+1} = 10.0 \times x_1^t + a \times y_1^t + b \times y_3^t, \tag{2}$$
$$y_2^{t+1} = 3.0 \times x_3^t + 1.0 \times y_2^t + c \times y_3^t, \tag{3}$$
$$y_3^{t+1} = 2.0 \times x_2^t + 0.5 \times x_3^t + d \times y_3^t. \tag{4}$$

The task for subjects is first to explore the system (i.e. the causal links between the system variables) and then to control the endogenous variables (= the numbers of y-creatures) by means of the exogenous variables with respect to a set of given goal states. Parameters a-d are manipulated depending on the experimental conditions (see below).

2.2 General Experimental Procedure

In our experiments, subjects pass through at least two phases. In the first phase, the "*knowledge acquisition phase,*" subjects can explore the system and its behavior as they like (see also Moray, Lootsteen & Pajak, 1986; Shrager & Klahr, 1986). They can take actions (i.e., make an intervention on one or more of the exogenous variables) and observe the resulting effects in the endogenous variables. Figure 2 shows how SINUS is presented to subjects.

SINUS	Block 1				
Week	1	2	3	4	5
State:					
Gaseln.....	1600	1700	1800	1900	2000
Schmorken..	900	957	1013	1055	1096
Sisen......	300	293	286	281	306
Intervention:					
Olschen....	10	10	10	10	_?
Mukern.....	12	11	13	28	?
Raskeln....	-1	-1	-5	-5	?

Press "space bar" to select an intervention, choose a value and then press "return"

Figure 2. Screen display of DYNAMIS when presenting system SINUS after 4 weeks (= trials) on the first block. The upper part shows the state of the three endogenous variables, the lower part shows past interventions.

Each block consists of a certain number of trials (referred to as "weeks" in the cover story) which all depend on each other. From one block to another the system is reset to the same starting values. From time to time we measure the knowledge that has been acquired so far by asking the subjects for a graphical representation of their structural knowledge ("causal diagrams"). In the second phase, called "*knowledge application phase*," the subject has to reach a defined system state and try to maintain the variable values as close as possible to the values defined as goal states. In this phase, we measure the quality of the operator's control by assessing the distance between the current and the goal values for all endogenous variables. Some comments on measuring structural knowledge and system performance seem necessary at this point because this is central to our studies. A review on techniques for knowledge assessment can be found in Kluwe (1988). Also, Rouse and Morris (1986) discuss some of the diagnostic problems in more detail.

2.3 Measuring Structural Knowledge and System Performance[3]

Starting with *control performance quality*, the goal is to determine how well a given goal state is approximated by the operator's interventions. The classical approach requires the measurement of the deviation from the target system state in terms of the root mean squares criterion (RMS). This indicator reflects the mean deviation, independent of sign. The weights of individual

3 This section follows the presentation as given in Funke (1991).

deviations become increasingly higher the farther away they are from the target state. A good discussion of the frequently used RMS criterion can be found in Poulton (1973) and Bösser (1983).

There is, however, an aspect which reflects an ugly property of this kind of system performance evaluation: Assume that an operator has knowledge about the system. Reaching the goal is of little or no difficulty for him and the resulting RMS will be low. But what can we say about the operator with little or no knowledge? The resulting distance to the goal state as measured by the RMS criterion varies as a function of his (random) interventions. Depending on the weight matrices of the system, this would result in a large variety of measured distances. Therefore, different values of the RMS, in this case, do not reflect different degrees of quality of system performance. The argument here is one of different reliabilities of the RMS criterion for different states of an operator's knowledge, being best in the case of correct knowledge (RMS indicating reliable values near zero) and worst in the case of purely random intervention (RMS indicating an enormous range of values due to decreasing reliability).

One potential solution for this problem is a *logarithmic transformation* of the RMS. This transformation leads to an evaluation of distances which is more efficient: Larger distances are no longer weighted more heavily. Rather, they are considered less important by this measure. It does not matter if someone missed the goal by 10,000 or 100,000 points. This difference has the same importance as the difference between a deviation of 1 and 10. The transformation, thus, reduces the error variance that increases as a function of the operator's distance to the goal state. In the experimental section the variable "QSC" refers to this kind of dependent variable ("Quality of System Control"—a low QSC reflecting a good score because of low discrepancies between goal values and the values subjects reached on the endogenous variables through their control behavior).

Measuring the *structural knowledge* an operator has acquired about a system requires also some kind of distance or similarity measurement. In this case the distance exists between the structural relations hypothesized by subjects and those implemented in the system. For this purpose, the operator marks on a sheet (or in some versions directly on the screen) the assumed causal relationships at certain points in time. The problems with this kind of measurement are: (1) *Response tendency*: Subjects differ in their degree with which they indicate relations in cases they are not quite sure about. Therefore, one has to count both hits (i.e., correspondence between assumed and existing relation) and false alarms (like in conventional recognition tests). (2) *Different quality of knowledge*: On the lowest level it is only assumed that a relationship between two variables exists (*relational knowledge*). On the next higher level the sign of the relation is known (*sign knowledge*). In the optimal case the numerical weights are known (*numerical knowledge*). (3) *Functionality instead of correctness*: False models can be useful for system control, at least within a restricted area of values. Similarity measures are blind to this aspect. (4) *Implicit assumptions*: Consider, for example, a variable that does not change over time. The weight for that relation is assumed to be equal to one. This kind of knowledge a subject often does not find worth mentioning. (5) *Single vs. multiple models*: It is not clear whether a subject follows only one single model at a given point in time or whether there exist several models concurrently.

For problems (3) to (5), no solution can be given at present. Problems (1) and (2), however, can be solved using a quantification of the following kind: For each causal specification of a subject one first counts whether it belongs to one of the three classes of knowledge (relational, sign, or numerical) and whether it is correct or false. Then, for each of the three levels one can determine the "quality of system identification" (QSI) in terms of the difference between "hits" (HI) and "false alarms" (FA), weighted by a "guessing" probability (p) according to the following scheme which closely resembles the discrimination index P_r from the two-high threshold model for recognition memory (see Snodgrass & Corwin, 1988; the proposed "correction for guessing" dates back to Woodworth, 1938):

$$QI = (1-p) \cdot \frac{HI}{\max(HI)} - p \cdot \frac{FA}{\max(FA)}, \qquad -p \leq QI \leq (1-p) \tag{5}$$

The guessing probability for numerical parameters in a dynamic system could, for instance, be set to zero. In this case all hits are counted relative to the maximal number of hits, max(HI). If one sets the guessing probability to 0.5 in the case of sign knowledge, then errors lead to a reduction in the QSI index for that level. The index for structural knowledge, which serves as a dependent variable in the following experiments, is called "QSI" ("Quality of Identification"—high QSI revealing a good score because of high correspondence between implemented and assumed causal relations) and results from an additive combination of the QSI-values for all three knowledge levels. A study by Müller (in press) demonstrates considerable reliability and, thus, sufficient psychometric quality of this index.

3. Experimental Studies on System Properties

In the following section five experiments on the role of different system properties serve to illustrate the approach just outlined. The focus of the experiments is on the role of active intervention into a system vs. pure observation (Exp. 1), on the influence of different degrees of Eigendynamik (Exp. 2) and side effects (Exp. 3), and on the effects of presentation mode, prior knowledge, controllability of the system, and degree of control required (Exp. 4 & 5). For each of the experiments the presentation includes a description of the independent as well as dependent variables, subjects, material, procedure, hypotheses, result, and a short discussion. Then, in a next section, a general discussion selects interesting results and connects them with results from other studies.

3.1 Experiment 1: Active Intervention vs. Pure Observation

Independent and dependent variables. In this first experiment (for more details see Funke & Müller, 1988) learning by active interventions was compared to learning by pure observation of the system's development (Factor 1: intervention vs. observation, I vs. O). In addition, the effect of a diagnostic tool (subjects had to predict the next system's state) was compared to a nonprediction condition (Factor 2: prediction vs. no prediction, P vs. NP). Dependent variables were QSC and QSI.

Subjects, material and procedure. Subjects were 32 college students. In each of the four conditions eight subjects were run individually. This allows for detection of "large effects" ($f = 0.40$ in Cohen's meaning of the word, 1977) with $\alpha = 0.10$ and $\beta = 0.30$ for main effects. In the I- and O-condition experimental twins were used. Each subject in the O-condition observed the system data which another subject (the twin) under the I-condition had produced (yoked-control design).

The system used was SINUS with parameters $a = 1$, $b = 0$, $c = 0.2$, and $d = 0.9$ in Eq. (2), (3) and (4). The system had to be manipulated during five blocks of seven trials each. During the first four blocks subjects could freely explore the system. During the fifth block all subjects (both the I- and the O-group) had to reach and maintain a previously specified goal state. The amount of system knowledge subjects had acquired (QSI, as measured by the "causal diagram" at the end of exploration) and the control quality (QSC, as measured via the distance of the actual to the specified goal states) served as dependent variables.

Results. Funke and Müller expected the I-group to be superior to the "observers" with regard to amount of knowledge as well as to control quality. Also, the "predictors" should accumulate more knowledge than the "non-predictors." Path-analytical evaluation of the data supported the expectations partially: The I-group was indeed better in controlling the system (significant standardized path coefficient $\beta = 0.42^*$ from I to QSC), but seemed to know less than the "observers" ($\beta = -0.30^*$ from I to QSI). "Predictors" had more knowledge than "non-predictors" (mean QSI: 1.02 vs. 0.57, $F_{(1,28)} = 5.50^*$). Knowledge about the system was generally a good predictor of control performance ($\beta = 0.41^*$ from QSI to QSC). Interestingly, there was a negative relationship between the time spent on the task and the quality of performance.

Discussion. The results demonstrate the effects of task manipulations. Active interventions allow better system control. However, this effect is not accompanied by an increase in "externalizable" knowledge. Similar dissociations have been reported by Broadbent, FitzGerald, and Broadbent (1986), Berry and Broadbent (1984, 1987), and Putz-Osterloh (1987), for a critique see Sanderson (1989). Requiring subjects to predict the next state increases the amount of knowledge as revealed by QSI. Detailed analyses of so-called "experimental twins"—pairs of subjects who had to cope with the same system situations—indicated a high interindividual variability: There were no significant correlations between the twins' QSI and QSC scores, thus showing the importance of person-specific ways of information processing.

3.2 Experiment 2: Effects of Eigendynamik

Independent and dependent variables. In this second experiment the effect of different degrees of "Eigendynamik" was analyzed. Eigendynamik means that an endogenous variable at time t has an effect on its own state at time t+1. Thus, parameters a and d from Figure 1 and Eq. (2) to (4) were changed in three steps: (1) $a = 1$, $d = 1$: a control condition without any Eigendynamik; (2) $a = 1$, $d = 0.9$: one variable with Eigendynamik; (3) $a = 1.1$, $d = 0.9$: two variables with Eigendynamik. Parameters $b = 0$ and $c = 0.2$ were held constant. Dependent variables were QSC and QSI.

Subjects, material, and procedure. A total of 24 paid males doing their civil service served as subjects. Under each of the three conditions eight subjects were run individually. Assuming $\alpha = 0.10$ and "large effects" ($f = 0.40$), the power $1\text{-}\beta$ proves to be at 0.50 in this case for the main effect (Cohen, 1977). SINUS was used to simulate the system with the characteristics described above. The system had to be manipulated during five blocks of seven trials each. During the first four blocks subjects could freely explore the system. During the fifth block all subjects had to reach and maintain a previously specified goal state.

Results. It was expected that with an increase in Eigendynamik the amount of acquired knowledge as well as the degree of control over the system should deteriorate. Analysis of variance revealed only a significant effect for QSC ($F_{(2,21)} = 3.23^{*}$; mean QSC for Eigendynamik of 0, 1 and 2 are 3.86, 3.70 and 5.18), but not for QSI ($F_{(2,21)} = 1.12$, n.s.).

Discussion. The results show that the degree of knowledge acquisition does not seem to be influenced by Eigendynamik. In contrast, the control of the system varied as a function of Eigendynamik. Particularly under the condition of two variables with Eigendynamik, control of the system turned out to be much more difficult. This points to the fact that knowledge acquisition and knowledge application require different abilities, which under certain circumstances lead to a dissociation of both measures.

3.3 Experiment 3: Identification of Side Effects

Independent and dependent variables. In this third experiment the effect of three different degrees of side effects was analyzed. Side effects were operationalized as minor effects from one endogenous variable on to another. In this case, parameters a and d from Figure 1 and Eq. (2) to (4) remained unchanged (1 resp. 0.9), but parameters b and c were changed in three steps: (1) $b = 0$, $c = 0$: a control condition without side effects; (2) $b = 0$, $c = 0.2$: one side effect; (3) $b = 0.5$, $c = 0.2$: two side effects. Dependent variables were QSC and QSI.

Subjects, material, and procedure. Under each of the three conditions eight male subjects were run individually. According to Cohen (1977), assuming $\alpha = 0.10$ and "large effects" ($f = 0.40$) power $1-\beta$ proves to be 0.47 for the main effect. The system used was again SINUS with the changes described above and with the following change of the procedure. During the first four blocks exploration was not limited by number of trials but by time (15 min per block). During the fifth block all subjects were required to reach and maintain the previously specified goal states over seven trials, but without time pressure.

Results. The expected influence of side effects on knowledge acquisition was confirmed by a significant negative path coefficient ($\beta = -0.35^*$) from the side effect predictor to QSI (mean QSI for 0, 1 and 2 side effects are 1.14, 1.26 and 0.77, $F_{(2,21)} = 1.74$, n.s., respectively). Also, the effect from knowledge onto control quality reached significance ($\beta = 0.73^*$ from QSI to QSC; mean QSC for 0, 1 and 2 side effects are 2.39, 2.86 and 4.72, $F_{(2,21)} = 4.01^*$, respectively). The number of trials in blocks 1 to 4 had (contrary to our expectation) no predictive value for QSC or QSI, but this conclusion is taken only as preliminary because of medium power.

Discussion. As in previous experiments, the manipulation of another system attribute shows an effect on knowledge acquisition as revealed by the QSI measure and, again, the amount of knowledge predicts the quality of system control. This result is in line with Conant and Ashby (1970). According to these authors, good control has to be a consequence of a good model.

3.4 Experiment 4 and 5: Effects of Presentation Mode, Prior Knowledge, Controllability and Amount of Control

Material. For the last two experiments a different system was used called ALTÖL (=used oil) which was designed to activate prior knowledge. Therefore, in a pilot study based on interviews with 32 students we assessed the relations between variables within the area of the "ecological load of used oil." A "modal" model was constructed using the relations which were named by at least 23 of the participants in the pilot study. The resulting relations are depicted in Figure 3.

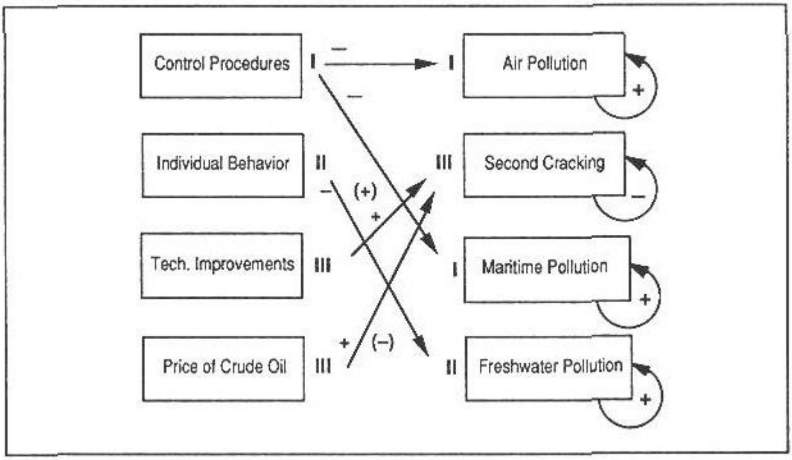

Figure 3. Causal structure of the ALTÖL simulation. The signs in parentheses are valid only in the "mismatching version" (see text). The Roman figures indicate the numbers of the partial systems which are mutually independent.

Independent and dependent variables. Experimental variations were made with respect to the following factors. (1) Knowledge compatibility: Two versions of ALTÖL were realized such that the signs of the parameters of two of the four exogenous effects were either in concordance with the assumed prior knowledge (matching condition) or disagreed with it (mismatching condition). The question was how the semantic embedding inhibits or facilitates the identification of two systems with identical structures. (2) In one condition subjects had to control all four endogenous variables on the last block, whereas a second condition required the control of only two of the variables. (3) The experiment also investigated whether control of the system would be influenced by the number of exogenous variables by which subjects could manipulate the endogenous variables. This was a within-subjects factor. The ALTÖL system consisted of three independent parts (marked by I, II and III in Figure 3) with a 1:2, 1:1 and 2:1 relation between exogenous and endogenous variables. (4) The last factor manipulated the user interface of the system: In one condition the variables were presented numerically (Exp. 4), the second condition used a graphical presentation which contained the same information (Exp. 5; see Fig. 4).

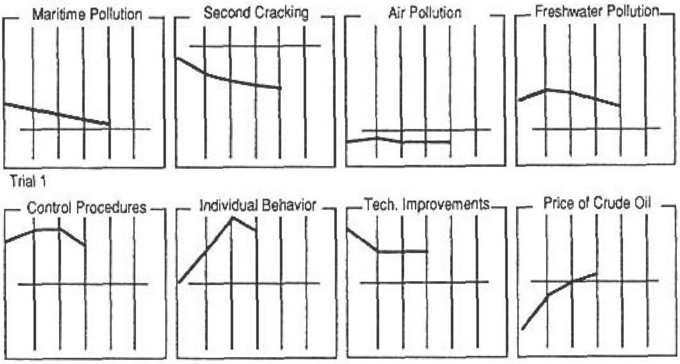

Figure 4. Screen display of ALTÖL on the fifth simulated week of a trial (graphical condition). Upper part: endogenous variables (straight lines: goal values), lower part: exogenous variables (straight lines: zero intervention). The end points of the exogenous variables can be dragged in the upper or lower edge of the box in order to indicate the degree of a positive or negative intervention.

Dependent variables were QSC and QSI. Because for subjects in the condition with only partial system control a general QSC-score could not be computed, a separate QSC is computed on the basis of the two partial systems these subjects had to control.

Subjects and procedure. A total of 80 Bonn University students served as subjects for both experiments, such that under each of the eight conditions 10 subjects were used. Assuming $\alpha = 0.10$ and "large effects" ($f = 0.40$) power $1\text{-}\beta$ proves to be at 0.96 for the main as well as for the interaction effects (Cohen, 1977). Subjects were either given a small honorarium or they fulfilled course requirements.

Results. With respect to knowledge compatibility it was expected that a system with counterintuitive relationships would diminish knowledge acquisition as well as control performance. On QSI, a significant main effect of knowledge compatibility was observed (mean QSI for matching and mismatching version is 0.31 and 0.17, $F_{(1,72)} = 3.92^{*}$, respectively). Also, control performance was affected (mean QSC for matching and mismatching version are 4.10 and 8.28, $F_{(1,36)} = 13.86^{*}$, respectively; note that only for subjects with four goals QSC could be computed).

A second hypothesis specified effects of presentation format: The graphical user interface of the system should facilitate the identification of system structure (QSI), whereas the numerical condition should facilitate control performance (QSC). Both expectations were disconfirmed. Neither QSI nor QSC showed significant main effects of presentation mode.

Presentation effects can be observed if one takes into consideration the third factor, the number of goals to be reached, for which a main effect was expected but did not occur on QSI ($F < 1$). The interaction between presentation mode and number of goals is significant (mean QSI for the numerical version is 0.38 and 0.14 for 2 and 4 goals, for the graphical version it is 0.08 and 0.36, $F_{(1,72)} = 13.36^*$). Given numerical presentation and only 2 goals the causal analysis of the system was better than for the graphical version. With 4 goals the effect was opposite, producing better identification under graphical than under numerical presentation.

A final hypothesis specified that a 2:1 relation of exogenous to endogenous variables (condition I, high degree of controllability) would result in better control than the reverse relation (condition III, low degree of controllability). Contrary to this expectation, subjects showed best system control under condition III, medium control under II, and worst control under I (mean QSC for I, II, and III are 4.75, 6.98, and 8.13; $F_{(2,72)} = 5.26^*$, respectively).

Discussion. It is not surprising that the match or mismatch between prior knowledge and implemented system structure shows strong effects on knowledge acquisition and control performance. This illustrates an effect which could implicitly occur in studies where subjects' prior knowledge remains uncontrolled. In addition to the classical methods of analyzing effects of prior knowledge by comparing novices and experts (Reither, 1981) or by comparing a semantically embedded system with an abstract one (Hesse, 1982), the method of constructing two semantically equivalent systems which correspond differently to prior knowledge proves to be useful.

Presentation format *per se* was not a critical factor. However, it is obvious that depending on the nature of the task, differential effects occur: In order to cope with the more complex task the graphical presentation which is less precise in presenting system information yields better results. It remains unclear why different degrees of controllability did not have their expected effect on control performance: Selective motivational effects (the task which appears more difficult motivates subjects to work harder on it during the control phase) could be one possible explanation for this surprising result.

4. General Discussion

The main conclusions of the five experiments presented above are as follows. First, a research program of manipulating formal system attributes experimentally contributes to an exploration of the differential influence of these attributes on knowledge acquisition and knowledge application. Second, the dependent variables for the amount of causal knowledge and precision of system control seem useful indicators in studies designed to analyze acquisition and application of knowledge. Third, the distinction between different kinds of relations within a system (endogenous and exogenous effects, Eigendynamik, side effects) should be further explored because QSI and QSC seem to be influenced by that kind. Fourth, the formal approach is not restricted to systems with "artificial" semantic embedding but is also useful for semantically rich systems. Fifth, besides variations of system attributes the effects of different presentation modes (active vs. passive exploration; numerical vs. graphical display) have also been found to be influential.

Still lacking is a theory of how people build up a mental representation of a system and how they derive interventions from that model. Such a theory should also explain under which circumstances failures in a system could be detected and how, for example, operators can cope with dangerous situations. Also, the model should incorporate human error which occurs even in cases where perfect knowledge is available (e.g., with overtrained pilots).

Klahr and Dunbar (1988) recently developed an integrated model for scientific reasoning that seems to be applicable to our scenarios. They argued, "that scientific reasoning can be conceptualized as a search through two problem spaces: an hypothesis space and an experiment space" (1988, p. 7). These spaces result from the task of a (naive or well-trained) scientist:

"The successful scientist, like the successful explorer, must master two related skills: knowing where to look and understanding what is seen. The first skill—experimental design—involves the design of experimental and observational procedures. The second skill—hypothesis formation—involves the formation and evaluation of theory." (Klahr & Dunbar, 1988, p. 2).

This assumption seems applicable to the research topic of this paper insofar as the experimental situation Klahr and Dunbar were concerned with—exploration of a hitherto unknown object—is basically identical to the situation of exploring and controlling an unknown linear equation system. Furthermore, they conceptualized the process of knowledge acquisition in terms of hypotheses that are developed either more inductively or more deductively. With respect to interindividual differences one could assume that some of the subjects follow more closely the hypothesis-oriented approach (and do some sort of model testing), whereas others proceed more data oriented (and try to inductively form a model).

Besides the great similarity in the experimental procedure and in the theoretical frame of reference, however, there are some differences. The main difference can be seen in the way subjects' knowledge is measured. Klahr and Dunbar primarily used verbal data (see Bainbridge, 1979, as well as Ericsson & Simon, 1980, for a critical comment on verbal data in this context), whereas in our procedure different approaches are taken to diagnose the structural knowledge a subject acquires. This difference is partly due to our "object" of exploration: Subjects explicitly have to anticipate the next states of the system, they have to write down their hypotheses about structural relationships, and they have to control the system as well as possible.

Brehmer (1989) conceptualizes process control in terms of "dynamic decision tasks"—in contrast to static or sequential decision tasks—with the following four characteristics: "(a) a series of decisions are required; (b) these decisions are interdependent; (c) the decision problem changes, both autonomously and as a consequence of the decision maker's action; and (d) the decisions are made in real time" (Brehmer, 1989, p. 144). With exception of the last criterion, this definition is also applicable to the systems SINUS and ALTÖL.

What are the main goals for future research? Three tasks will be outlined briefly: (1) a differentiation between factors influencing complex problem solving that come from individual, situational and system attributes; (2) reliability and validity research on complex problem solving scenarios; (3) adequate measurement of the actual "mental model" and of the potential heuristics that complex problem solvers use, also over time.

Concerning the first task, *separation of person, situation, and system influences* on performance measures, the approach taken by Streufert et al. (1988) seems to point into an interesting direction: Instead of using "free" simulations in which decisions can change a system's state quite drastically, they use a "quasi-experimental simulation technology," in which the system reacts in part independently of subjects' interventions such that each subject receives comparable information and events. Despite this fact, subjects still believe that they have a direct or delayed impact on the system. This technique should be explored further in order to standardize the conditions under which subjects' performance quality is measured independent of system attributes. Also, one has to make clear how this kind of pseudo-feedback affects knowledge acquisition or even prevents it. In the case of small-scale systems like SINUS the technique may not be indicated, but in large-scale systems like LOHHAUSEN the real feedback appears not to be of central importance. In this latter case, strange constellations may seem plausible because even in case of correct feedback not all the information can be processed by the problem solver. In the former case of a deterministic small-scale system, pseudo-feedback would quickly cause irritation.

Even if subjects did not detect different degrees of side effects or Eigendynamik—an argument used recently by Strohschneider (1991) in a critique of the present experimental approach—the effects of this manipulation on the dependent measures cannot be denied. Also, Brehmer and Allard (1991) point to the critical role of system characteristics. In their studies, introducing different degrees of feedback delay leads to a detrimental problem solving behavior (see also Sterman, 1989).

Concerning task 2, *reliability and validity aspects*, there is a good deal of work to be done, and it is not quite clear how to proceed: Up to now, researchers mainly pointed to the face validity of their tasks and dependent measures. Jäger (1986, p. 274) called this "uncovered checks" that have to be cashed in subsequent research. It is simply not enough to show that there are no correlations between dynamic tasks and standard intelligence tests because many reasons can account for such results. Rather, one has to show positive connections to other psychometric instruments as well as to external criteria. Reliability studies have to demonstrate that the indices used do not show "random walk" characteristics. Much progress would also be possible if at least results from empirical studies were subjected to replication. By now, such replications are widely missing. If no direct assessment of reliability seems possible, then at least replications should be presented in order to establish stable phenomena.

Traditional measurement theory is not the only way to give answers to this question. Modern concepts like latent state-trait theory (see Steyer, Widaman, & Gräser, in press), with their distinction between consistency and specificity of certain variables, offer the potential for careful design of reliability studies such as the one conducted by Müller (in press) which revealed good scores for the QSC and QSI measures but also showed the existence of situational influences.

One possible line of validity research could be the use of the learning test concept (e.g., Guthke, 1982) according to which intelligence is not a static variable but must be interpreted as "learning potential." It could be possible that relations exist between this "learning potential" and the ability to solve complex problems. Following a solid comparison of the two different requirements, predictions are possible concerning what kind of selected dependent variables from both areas should correlate.

Concerning task 3, the *adequate measuring of the problem solvers' mental models and their heuristics*, one has to develop instruments that sensitively assess those relevant parts of human memory that are required for identification and control. Whereas in the area of assessing structural (or declarative) knowledge some useful techniques exist, there are clear deficits in diagnosing the heuristic knowledge that human problem solvers operate with. Also, more attention should be given to developing measurement techniques that reveal the implicit knowledge of an operator. These could be developed similar to the procedures used in experimental memory research: If the recent use of a representation or of a procedure facilitates its subsequent access (as revealed, for instance, by reaction time measurements in verification tasks on these representations or procedures), then indirect effects could be established in roughly the same manner as in research on human memory.

Reaction time measurements have proven useful in the area of process control of finite state automata (see Funke & Buchner, in press). Based on a formal analysis of the task under study it was demonstrated that a facilitating effect on RT occurs in verification situations in which the "natural" sequence of state transitions is reinstated during test.

Concerning the *general research strategy*, I find it more useful to manipulate critical variables in system structures and in presentation modes than to create numerous new systems which are completely unrelated and offer no solid basis for comparisons. Collecting data without theoretical assumptions produces puzzling situations in which spurious correlations may suggest significant effects where no effects are present. Only the strategy of analyzing the effects of selected variations based on some minimal theoretical premises—the experimental method—can offer new insights into the principles and mechanisms that govern complex human problem solving. For this purpose, the research strategy outlined above offers a method for the systematic construction and variation of stimulus material with well known characteristics to be used in future experiments.

5. References

Bainbridge, L. (1979). Verbal reports as evidence of the process operator's knowledge. *International Journal of Man-Machine Studies, 11*, 411–436.

Berry, D. C., & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology, 36*, 209–231.

Berry, D. C., & Broadbent, D. E. (1987). The combination of explicit and implicit learning processes in task control. *Psychological Research, 49*, 7–15.

Bhaskar, R., & Simon, H. A. (1977). Problem solving in semantically rich domains: An example from engineering thermodynamics. *Cognitive Science, 1*, 193–215.

Bösser, T. (1983). Eine nichtlineare Regelstrategie bei der manuellen Regelung [A nonlinear control strategy at manual control]. *Zeitschrift für Experimentelle und Angewandte Psychologie, 30*, 529–565.

Brehmer, B. (1989). Dynamic decision making. In A. P. Sage (Ed.), *Concise encyclopedia of information processing in systems and organizations* (pp. 144–149). New York: Pergamon Press.

Brehmer, B., & Allard, R. (1991). Dynamic decision making: The effects of task complexity and feedback delay. In J. Rasmussen, B. Brehmer & J. Leplat (Eds.), *Distributed decision making: Towards a basis for modeling* (pp. 319–334). Chichester: Wiley.

Broadbent, D. E., & Aston, B. (1978). Human control of a simulated economic system. *Ergonomics, 21*, 1035–1043.

Broadbent, D. E., FitzGerald, P., & Broadbent, M. H. P. (1986). Implicit and explicit knowledge in the control of complex systems. *British Journal of Psychology, 77*, 33–50.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences (2nd ed.)*. New York: Academic Press.

Conant, R. C., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of System Science, 1*, 89–97.

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* (pp. 97–109). Chichester: Wiley.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*, 215–251.

Funke, J. (1984). Diagnose der westdeutschen Problemlöseforschung in Form einiger Thesen [Diagnosis of problem solving research in the FRG in the form of some theses]. *Sprache & Kognition, 3*, 159–171.

Funke, J. (1988). Using simulation to study complex problem solving. A review of studies in the FRG. *Simulation & Games, 19,* 277–303.

Funke, J. (1991). Solving complex problems: Exploration and control of complex systems. In R. J. Sternberg & P. A. Frensch (Eds.), *Complex problem solving: Principles and mechanisms* (pp. 185–222). Hillsdale, N. J.: Lawrence Erlbaum.

Funke, J., & Buchner, A. (in press). Finite Automaten als Instrumente für die Analyse von wissensgeleiteten Problemlöseprozessen: Vorstellung eines neuen Untersuchungsparadigmas [Finite automata as instruments for analyzing knowledge-based problem solving processes: Introduction of a new research paradigm]. *Sprache & Kognition, 10.*

Funke, J., & Müller, H. (1988). Eingreifen und Prognostizieren als Determinanten von Systemidentifikation und Systemsteuerung [Intervention and prediction as determinants of system identification and system control]. *Sprache & Kognition, 7,* 176–186.

Greeno, J. G. (1974). Hobbits and orcs: Acquisition of a sequential concept. *Cognitive Psychology, 6,* 270–292.

Guthke, J. (1982). The learning test concept—an alternative to the traditional static intelligence test. *The German Journal of Psychology, 6,* 306–324.

Hays, R. T., & Singer, M. J. (1989). *Simulation fidelity in training system design. Bridging the gap between reality and training.* New York: Springer.

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

Hübner, R. (1989). Methoden zur Analyse und Konstruktion von Aufgaben zur kognitiven Steuerung dynamischer Systeme [Methods for analysis and construction of tasks for cognitive control of dynamic systems]. *Zeitschrift für Experimentelle und Angewandte Psychologie, 36,* 271–238.

Jäger, A. O. (1986). Validität von Intelligenztests [Validity of intelligence tests]. *Diagnostica, 32,* 272–289.

Jeffries, R., Polson, P. G., & Razran, L. (1977). A process model for missionaries—cannibals and other river-crossing problems. *Cognitive Psychology, 9,* 412–440.

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

Kluwe, R. H. (1988). Methoden der Psychologie zur Gewinnung von Daten über menschliches Wissen [Psychological methods for data acquisition about human knowledge]. In H. Mandl & H. Spada (Eds.), *Wissenspsychologie* (pp. 359–385). München, FRG: Psychologie Verlags Union.

Mackinnon, A. J., & Wearing, A. J. (1985). Systems analysis and dynamic decision making. *Acta Psychologica, 58,* 159–172.

Moray, N., Lootsteen, P., & Pajak, J. (1986). Acquisition of process control skills. IEEE Transactions on Systems. *Man, & Cybernetics, 16,* 497–504.

Müller, H. (in press). Complex problem solving: The evaluation of reiiability, stability, and some causal models. In R. Steyer, H. Gräser & K. F. Widaman (Eds.), *Consistency and specificity: Latent state-trait models in differential psychology.* Heidelberg: Springer-Verlag.

Poulton, E. C. (1973). *Tracking skill and manual control.* New York: Academic Press.

Putz-Osterloh, W. (1987). Gibt es Experten für komplexe Probleme [Are there experts for complex problems]? *Zeitschrift für Psychologie, 195,* 63–84.

Putz-Osterloh, W., & Lemme, M. (1987). Knowledge and its intelligent application to problem solving. *The German Journal of Psychology, 11,* 286–303.

Reither, F. (1981). About thinking and acting of experts in complex situations. *Simulation & Games, 12,* 125–140.

Rouse, W. B., & Morris, N. M. (1986). On looking into the black box: Prospects and limits in the search for mental models. *Psychological Bulletin, 100,* 349–363.

Sanderson, P. M. (1989). Verbalizable knowledge and skilled task performance: Association, dissociation, and mental models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 729–747.

Shrager, J., & Klahr, D. (1986). Instructionless learning about a complex device. *International Journal of Man-Machine Learning, 25,* 153–189.

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

Sterman, J. D. (1989). Misperception of feedback in dynamic decision making. *Organizational Behavior and Human Decision Processes, 43,* 301–335.

Steyer, R. (1984). Causal linear stochastic dependencies: An introduction. In J. R. Nesselroade & A. von Eye (Eds.), *Individual development and social change: Explanatory analysis* (pp. 95–124). New York: Academic Press.

Steyer, R., Gräser, H., & Widaman, K. F. (Eds.). (in press). *Consistency and specificity: Latent state-trait models in differential psychology.* Heidelberg, FRG: Springer.

Streufert, S., Nogami, G. Y., Swezey, R. W., Pogash, R. M., & Piasecki, M. T. (1988). Computer assisted training of complex managerial performance. *Computers in Human Behavior, 4,* 77–88.

Strohschneider, S. (1991). Kein System von Systemen! Kommentar zu dem Aufsatz "Systemmerkmale als Determinanten des Umgangs mit dynamischen Systemen" von Joachim Funke [No system of systems! Comment on the paper "System characteristics as determinants of dealing with dynamic systems" by Joachim Funke]. *Sprache & Kognition, 10,* 109-113.

Woodworth, R. S. (1938). *Experimental psychology.* New York: Holt.

Joachim Funke, Psychologisches Institut, Universität Bonn, Römerstr. 164, D-5300 Bonn 1 (FRG).