An Explorative Study of Search of Model Space in Problem Solving

Saskia Kistner (kistner@paed.psych.uni-frankfurt.de)
Institute of Psychology, Goethe University, Grüneburgplatz 1, D-60323 Frankfurt/Main, Germany

Bruce D. Burns (bruce.burns@sydney.edu.au)
School of Psychology, University of Sydney, Brennan MacCallum Bldg, A18, Sydney, NSW 2006, Australia

Regina Vollmeyer (r.vollmeyer@paed.psych.uni-frankfurt.de)
Institute of Psychology, Goethe University, Grüneburgplatz 1, D-60323 Frankfurt/Main, Germany

Ulrich Kortenkamp (ulrich.kortenkamp@mathematik.uni-halle.de)
Institute of Mathematics, Martin-Luther-University, Theodor-Lieser-Straße 5, D-06120 Halle (Saale), Germany

Abstract

Building on dual-space theories, the three-space theory of problem solving suggests to add search of a model space in addition to search of experiment and hypothesis space. This study aimed at exploring the three postulated spaces, especially model space, by means of verbal protocols. Participants (n=32) were asked to think aloud while working with a computer based learning program. With this program they could learn about torques in physics using interactive graphics in which experiments with levers and forces could be conducted. Their knowledge about torques was tested before and after working with the program. Verbal protocols were analyzed with regard to the amount of search of the three spaces and regarding the quality of the participants' models for torques. The three postulated spaces could be reliably identified in the protocols. For validity, the model quality score was related to performance and predicted final knowledge beyond prior knowledge. Our results add to the validity of model space and allow us to derive more specific hypotheses from the three-space theory.

Keywords: learning in physics; problem solving; verbal protocols.

Introduction

Problem Solving as Search of Spaces

Problem solving, as commonly defined, occurs when an actual state has to be transformed to a goal state using appropriate operators. Since Newell and Simon (1972) problem solving is described as search of a problem space. This space contains possible states, which are searched to find a state that corresponds to the solution (goal state). In doing so, operators (e.g., inferences) are used to move between states. However, when considering problems that involve the induction of rules or the formation of new concepts, a single problem space does not seem appropriate anymore to describe phenomena comprehensively. In Greeno's (1978) typology of problems, which distinguishes problems of inducing structure, of transformation, and of arrangement, these kinds of problems could be assigned to the category of induction problems. To account for a larger scope of problems, including induction problems, the problem space framework was expanded to dual-space theories.

Simon and Lea (1974) suggested two spaces, a rule space and an instance space. The rule space contains rules that can be applied to the problem solving task, while the instance space consists of possible states of the problem. Rules that are generated in rule space are tested in instance space by applying them to instances. In turn, results of this test process are used as basis to modify the rules if necessary. Assuming two spaces, the problem space framework could not only be applied to solving simple problems like those studied by Newell and Simon (1972), but also to problems that involve rule induction and were until then studied from a concept formation view (e.g., Bruner, Goodnow, & Austin, 1956). While one problem space is sufficient to describe typical problem solving, rule induction needs both, a rule space and an instance space. Simon and Lea's (1974) dual-space theory is still regarded as useful in contemporary research on complex problem solving in computer-simulated scenarios (Fischer, Greiff, & Funke, 2012).

The idea of search of two spaces was expanded by Klahr and Dunbar (1988) in their Scientific Discovery as Dual Search (SDDS) theory. They focused on the process of scientific discovery, during which hypotheses have to be formed and experiments have to be designed and conducted. Similarly to induction problems, in scientific reasoning rules or relationships applying to the concepts under investigation have to be discovered. Klahr and Dunbar (1988) conceptualize scientific reasoning as search of two interacting problem spaces, similar to Simon and Lea's (1974) rule and instance space. Hypothesis space consists of generated hypotheses, and experiment space contains possible experiments that can be run. Search of hypothesis space is determined by prior knowledge on the one hand and by results from running experiment space on the other hand. Movement in experiment space, that is choosing and conducting experiments, is guided by the current hypothesis to be tested. Results of the experiments are evaluated again in hypothesis space.

Representation of the Problem

Greeno (1978) argues that one of the main processes involved in induction problems is understanding. So how is understanding represented in problem space search theories so far? In Newell and Simon's (1972) theory, the problem space consists of an internal representation of the problem, constructed by the problem solver. This representation could be considered as the problem solver's understanding of the problem. It does not need to be the same for every problem solver but depends on how the problem is understood. Hayes and Simon (1974) further elaborated the understanding process that takes place when a problem space is generated. If problem solving fails, the problem space has to be revised, which can involve a different understanding of the problem. In dual-space theories the understanding process got a bit out of sight, at least it was not explicitly referred to it. Klahr and Dunbar (1988) argue that prior knowledge should always be taken into account when studying scientific discovery, as in real scientific contexts prior knowledge is always relevant. They consider the possibility that the formulation of hypotheses is influenced by prior knowledge, especially in the beginning of working on a problem, when a frame for the problem is generated. In terms of Newell and Simon, prior knowledge could affect how the problem is internally represented, that is what the problem space looks like.

The Three-Space-Search Theory of Problem Solving

Evidence for the importance to consider the impact of different initial representations of a problem was found by Burns and Vollmeyer (2002) in their research on dual-space search. They analyzed verbal protocols of participants who had to discover the links between inputs and outputs in a linear system. Participants started with very different hypotheses which represented different ideas of what kinds of links could be considered. Thus, learners seemed to have a certain model of the task which determined the hypotheses that they took into account. The term *model* could be translated as the current representation or understanding of the task to be worked with or the concept to be learned. Their results of verbal protocol analyses led Burns and Vollmeyer (2000) to suggest a third space, model space, which consists of possible representations of a problem or a concept. The current state in model space (the current representation) defines the hypothesis space as it determines the possible hypotheses to be tested. Hypothesis space in turn interacts with experiment space. Experiments are conducted to test the hypotheses and results are evaluated to confirm or reject hypotheses. When hypotheses are (repeatedly) rejected by results of the experiments, there may come a point when movement in model space is necessary and model space is searched for another representation that leads to the formulation of different hypotheses.

Aim of the Study

In the present study we attempt to further investigate the three-space theory outlined above, especially to explore the concept of model space. The task we use to study the three spaces is a computer based physics learning program on the concept of torques. In this program, learners are presented with interactive graphics where they can manipulate levers and forces and observe the effects. This design allows them to conduct experiments by manipulating the graphics and so to test hypotheses about relations between levers, forces, and torques. Thus, the task can be seen as to build a concept of how torques work, which includes induction of rules that apply to this concept. Rather than other tasks that are often used when studying problem solving, concept formation or rule induction, this task resembles to a greater extent learning tasks that students are faced with in school. While participants were working with the computer program, verbal protocols were recorded.

This study set out to answer three questions: (1) Can we reliably distinguish search of model space from search of hypothesis or experiment space in the verbal protocols? (2) Can we show the validity of our measures for search of model space? (3) Can search of model space predict final knowledge better than search of hypothesis space?

Method

The Problem Solving Task

The problem solving task in this study was the computer based learning program "How to visualize torques" (Wünscher & Ehmke, 2002). Working with this program, students can acquire an understanding of the concept of torques and learn about variables that can be considered for torques, e.g. the length or shape of a lever, or the power or angle of a force. The program consists of five units on different aspects of torques. It includes twelve interactive graphics in which students can manipulate levers and forces.

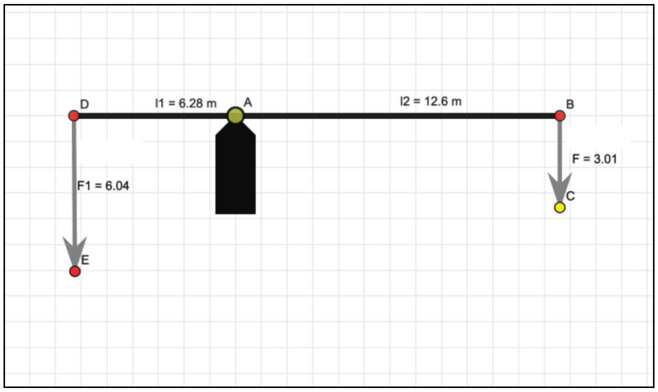

Figure 1: Example of an interactive graphic in the computer based learning program (red points can be moved).

In the graphic shown in Figure 1 for example, they can increase or decrease the length of the levers and/or the power of a force. As they do so, the other variables adjust to keep the construction balanced and the students can observe the effects of their manipulation. The program is composed in a way that at first students are given the opportunity to discover certain principles of torques on their own by working with an interactive graphic. Afterwards they are presented with a text containing information and explanations.

Procedure

Participants were 32 undergraduate psychology students (mean age: 25.3 years, *SD* = 9.7, 72% female). Their prior knowledge on torques was assessed by a test (prior knowledge test). Then they received a standardized instruction to think aloud including some warm-up tasks in order to facilitate verbalization (Ericsson & Simon, 1993). In the warm-up tasks participants were asked to think aloud while multiplying two numbers in the head, while counting the windows in their parents' home and while naming 20 animals. Students had then 35 minutes to work with the learning program and were asked to think aloud while doing this. After learning with the program, participants were given another test to assess their final knowledge on torques (final knowledge test). All together the procedure took about 90 minutes.

Knowledge Tests

The final knowledge test contained twelve items on torques. The items were of different difficulties and of various formats. For some items students had to calculate, in others they had to draw or to give written answers. The test included items whose answers could be directly learned from the program, but also transfer items. Students could reach up to 64 points in this test. The prior knowledge test consisted of five items that were taken from the final knowledge test and were representative of the diverse difficulties and formats. Here the maximum score was 21.

Analysis of Verbal Protocols

Coding search of the three spaces To analyze the verbal protocols, we developed a coding system, using sentences as coding segments. Each sentence was coded as referring to experiment space, hypothesis space or model space. Experiment space was coded if participants verbalized their manipulations of the graphics. Hypothesis space was coded if participants stated hypotheses or if they formulated simple rules that they derived from experimenting with the graphics. Model space was coded if verbalizations showed that participants had an understanding of the concept of torques or related concepts. Indicators for understanding were for example participants formulating explanations for observed phenomena or drawing analogies. For coding model space it did not matter whether the participants' understanding of torques was correct or wrong from an objective point of view. For verbalizations that did not refer to the three spaces, we had different categories such as metacognition or motivation and emotion, which are not further described here. Table 1 gives examples for statements that were coded as verbalizations of search of the three spaces. Interraterreliability was computed for four of the 32 protocols. We got a mean Cohen's kappa of .72 which can be considered as substantial (Landis & Koch, 1977).

Analysis of the protocols resulted in a number of sentences referring to each of the three spaces. To account for different total lengths of the protocols, we also calculated the percentage of each category in the whole protocol by dividing the number of sentences of the category by the total number of sentences in the protocol.

Table 1: Example statements for categories of the coding system.

Category	Example statement
Experiment space	Let's see what happens if I pull this lever … ah, then F gets greater, the force increases to 10.
Hypothesis space	I guess if I double l2, F1 will presumably get half as great.
Model space	If I pull with a greater force, the lever must logically become shorter to keep the balance.

Rating quality of the models (model score) In addition to the amount of sentences coded as model space, it seemed also important what kind of a model participants had, that is what their understanding of torques was like. So we also rated the quality of the model statements. Consecutive sentences whose content dealt with the same aspect of torques were merged and counted as one model. Each of these models was then rated with regard to quality by assigning a score of 0, 1 or 2 (correctness score). Here, correctness and precision were taken into account. The model was assigned 0 if someone stated an idea that was completely incorrect. A score of 1 was assigned for partially correct or imprecise models, and 2 was assigned for correct and precisely stated models. Additionally to the correctness score, each model could get an extra point in either two ways. First, a quality rating of a model should also take into account whether participants developed the model on their own or if they learned a new principle of torques from the explanations presented in the text. Thus, we scored one extra point when a model was self-generated by the participants while working with a graphic (self-generation score). Second, the ability to draw analogies can be seen as an indicator of an elaborated understanding. Thus, models that included analogies were scored with an extra point (analogy score). A model could only get one extra point for either self-generation by participant or analogy. If both applied, analogy was given priority. The correctness score and either the self-generation or the analogy score were summed up for each model, so the maximum score a model could get was 3. Finally, if a participant's verbalizations included a model that was not, or only marginally, related to torques, the score of this model was halved.

The scores were summed up for all models in a participant's protocol, so that the total score (model score) represented quantity of search of model space as well as quality of the models. Every protocol was scored independently by two raters, who had to come to an agreement subsequently by discussion.

Results

Performance in the Knowledge Tests

In the prior knowledge test participants on average had a score of 5.4 (*SD* = 5.6) out of the maximum of 21. In the final knowledge test the average score was 29.4 (*SD* = 12.2) out of the maximum of 64. These two scores cannot be compared directly, as the prior knowledge test contained only five of the twelve final knowledge test items. To clearly demonstrate that learning took place in the task, another score was composed of the five items in the final knowledge test that were identical to the prior knowledge test. A mean score of 13.9 (*SD* = 4.7) resulted, which differed significantly from the prior knowledge test, *t*(31) = 8.4, *p* < .01.

Progress during the Learning Program

On average participants worked on 7.1 of the twelve interactive graphics (*SD* = 2.4). To learn everything that was required to answer the final knowledge test it was necessary to finish the eighth graphic. This graphic was finished by 21 participants. The last set of graphics was reached by 10 participants and finished by 2 of them.

Coding Search of the Three Spaces

Aim of the study was to identify the three postulated search spaces in the verbal protocols. On average the verbal protocols consisted of 144 (*SD* = 50.1) sentences. Table 2 gives an overview of the results of the coding procedure. The numbers of sentences coded as search of each of the spaces are reported as well as the percentage of each space in the whole protocols. For both measures, experiment space and model space were the most frequently coded spaces, whereas hypothesis space was found less often.

Table 2: Descriptive statistics of coding the three spaces in the verbal protocols.

	Number of sentences		Percentage in whole protocol	
	M	*SD*	*M*	*SD*
Experiment space	25.3	13.4	18.7	10.4
Hypothesis space	14.8	9.8	10.7	7.5
Model space	25.4	15.2	18.4	10.0

To see how the three spaces are related to each other, we looked at the intercorrelations. Regarding the numbers of sentences, hypothesis space and model space were correlated positively (*r* = .43, *p* = .02), whereas there were no significant correlations between the other spaces (experiment space and hypothesis space: *r* = -.07, *p* = .70; experiment space and model space: *r* = .21, *p* = .25). When looking at the percentage of sentences in the whole protocol, the pattern was different. The correlation between hypothesis space and model space was not significant (*r* = .13, *p* = .48), as well as the correlation between experiment space and model space (*r* = .09, *p* = .62). But here we found a significantly negative correlation between experiment space and hypothesis space (*r* = -.36, *p* = .04). We will later consider which of these measures is the better predictor for performance.

Rating Quality of the Models

The number of models per verbal protocol ranged from 1 to 26 (*M* = 9.6, *SD* = 5.3). The rating of the models resulted in a model score that ranged from 0 to 36.5 (*M* = 13.6, *SD* = 9.0). As stated in the method section, the model score reflects quality of the models as well as quantity of search of model space, as it was a sum of the scores of each of the participant's models. To have a pure measure of quality we also computed the mean score per model for every participant, which ranged from 0 to 2.3 (*M* = 1.3, *SD* = 0.6).

The different measures of model space are all interrelated (see Table 3). Correlations between measures that include quantity and quality ratings are higher than those with the pure quality measure (mean score per model). The model score shows the highest correlations with other measures.

Model Space and Performance

It was hypothesized that model space would be positively related to performance. Table 3 shows the correlations between the different measures of model space, the prior knowledge test and the final knowledge test. Most of the model space measures, except the percentage of model space, are significantly correlated with final knowledge. Among these measures, the model score shows the highest correlation with final knowledge. Though, most of the differences between the correlations are not significant. Only the correlation between percentage of model space and final knowledge differs significantly from the one between model score and final knowledge (*Z* = 2.03, *p* = .02). The model score and the mean score per model are also significantly related to prior knowledge. This could be expected from the three-space theory, as prior knowledge is assumed to influence the goodness of one's models.

When comparing the two measures number of sentences in model space and percentage of model space, we found that number of sentences is significantly correlated with final knowledge, whereas the percentage is not (Table 3). In this context, it should be considered that the total number of sentences in the verbal protocols almost significantly correlated with final knowledge, *r* = .31, *p* = .08.

Table 3: Correlations between measures of model space and performance.

	2	3	4	5	6	7
Number of sentences in model space (1)	.79**	.81**	.81**	.45**	.27	.52**
Percentage of model space (2)		.68**	.65**	.41*	.26	.34
Number of models (3)			.90**	.42*	.24	.66**
Model score (4)				.71**	.37*	.71**
Mean score per model (5)					.43*	.47**
Prior knowledge (6)						.53**
Final knowledge (7)						

**p<.01, *p<.05

Furthermore, model space was expected to predict final knowledge beyond prior knowledge. To test this, we ran a stepwise regression analysis with the final knowledge test as dependent variable, entering all five measures of model space as predictors, to identify the best predictor of final knowledge. Model score was extracted as the predictor that explained final knowledge best ($\beta = .71$, $p < .01$; $R^2 = .51$, $p < .01$). The other model space measures did not contribute incrementally to the prediction of final knowledge.

Having established model score as the best predictor, we chose it as the predictor for the regression analysis run to test the hypothesis that model space predicts final knowledge beyond prior knowledge. This analysis was run with the final knowledge test as dependent variable and the prior knowledge test as a predictor. When entering the model score as an additional predictor, R^2 increased significantly by .31 (Table 4).

Table 4: Regression analysis with final knowledge test as dependent variable.

Variable	β	R^2	ΔR^2
Step 1			
Prior knowledge	.53**	.28**	
Step 2			
Prior knowledge	.31*		
Model score	.60**		
		.59**	.31**

**p<.01, *p<.05

Components of the Model Score

As reported in the method section, the model score was composed of different components. Models were scored for correctness/precision and extra points could be assigned for self-generation of models by the participants or for drawing analogies. By correlating these components with final knowledge we intended to examine which of the components was most predictive of performance. The correlations with final knowledge as well as the intercorrelations between the model score components and model score are given in Table 5. The correctness component shows the highest correlation with final knowledge, followed by the self-generation component. The analogy component is not significantly related to final knowledge. The correctness score is also the component that correlates highest with the model score. These results indicate that considering self-generation of models and drawing analogies does not improve the model quality rating.

Table 5: Correlations between components of the model score and final knowledge

	2	3	4	5
Score correctness (1)	.77**	.54**	.99**	.73**
Score self-generation (2)		.30	.85**	.55**
Score analogy (3)			.57**	.31
Model score (4)				.71**
Final knowledge (5)				

**p<.01, *p<.05

Comparison of the Three Spaces

Did the three spaces differ in how predictive they are for performance? Table 6 shows the correlations between the three spaces (measured as number of sentences and percentage) and the final knowledge test. Experiment space is not significantly related to final knowledge in either of the measures. Hypothesis space correlates significantly with final knowledge when measured in number of sentences. Model space (number of sentences) shows the highest correlation with final knowledge, though compared to the correlation with hypothesis space (number of sentences) the difference is not significant ($Z = 0.63$, $p = .27$).

Table 6: Correlations between search of the three spaces and final knowledge

	Prior knowledge	Final knowledge
Number of sentences in experiment space	-.15	.07
Number of sentences in hypothesis space	.10	.39*
Number of sentences in model space	.27	.52**
Percentage of experiment space	-.19	-.17
Percentage of hypothesis space	.03	.23
Percentage of model space	.26	.34

**p<.01, *p<.05

Regarding prior knowledge, we did not find any significant correlations with the numbers and percentages of the three spaces (Table 6). With respect to model space these results contrast to the finding that the model score and the mean score per model, which reflect also the model quality, are significantly related to prior knowledge (Table 3). So, compared to hypothesis space and experiment space, for model space we found a relation to prior knowledge, at least with quality measures.

Discussion

This study set out to explore the concept of model space by examining three questions. How successful were we?

Can we reliably distinguish search of model space from search of hypothesis or experiment space?

Our analyses of the verbal protocols demonstrated that we were able to assign each sentence referring to knowledge acquisition to either experiment, hypothesis or model space. We could do this with acceptable reliability. The new construct of a model space had as many sentences as the experiment space and more than hypothesis space. In addition, we not only performed a quantitative coding, we also considered the total number of sentences the students produced while learning as well as the quality of the models. These different measures were moderately correlated. Given that model score correlates highly with all other scores we will use this measure for validation. Moreover, theoretically it expresses the goodness of the students' models.

Can we show the validity of the model score?

Our assumption was that having a good model for torques should help further knowledge acquisition. Therefore, model score should correlate with final knowledge even if prior knowledge is controlled. In a regression analysis we could demonstrate that indeed model score can predict final knowledge which is a clear validation of our coding system.

Can search of model space predict final knowledge better than search of hypothesis space?

Although the correlation between final knowledge and the number of sentences in model space is higher than the one with number of sentences in hypothesis space, the difference was not significant. This result was in the right direction, but better evidence for distinguishing between these spaces may come from the evidence of a higher correlation of prior knowledge with model score than with search of hypothesis space. Such a finding is consistent with the three-space theory in which the source of the initial model is prior knowledge. However, this answer needs caution and more research.

Future directions

Our long-term goal is to test whether the three-space theory can better predict complex problem solving or science learning than a two-space theory. This study took a step towards this goal by using verbal protocols to identify the types of models people formulate for torques and showing that the quality of such models relates to knowledge acquisition. The next step will be to develop an instrument to measure model quality in the domain of torques. Such an instrument will enable us to test further hypotheses derived from the three-space theory.

Acknowledgements

This research was financially supported by the Deutsche Forschungsgemeinschaft (VO 514/17-1).

References

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

Burns, B. D., & Vollmeyer, R. (2000). Problem solving: Phenomena in search of a thesis. In L. Gleitman & A. K. Joshi (Eds.), *Proceedings of the twenty-second annual meeting of the cognitive science society* (pp. 627–632). Hillsdale, NJ: Lawrence Erlbaum.

Burns, B. D., & Vollmeyer, R. (2002). Goal specificity effects on hypothesis testing in problem solving. *The Quarterly Journal of Experimental Psychology*, *55A*(1), 241–261.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (2nd ed.). Cambridge, MA: MIT Press.

Fischer, A., Greiff, S., & Funke, J. (2012). The process of solving complex problems. *The Journal of Problem Solving*, *4*(1), 19–42.

Greeno, J. G. (1978). Natures of problem-solving abilities. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes: Vol. 5. Human Information Processing*. Hillsdale, NJ: Erlbaum.

Hayes, J. R., & Simon, H. A. (1974). Understanding written problem instructions. In L. W. Gregg (Ed.), *Knowledge and cognition*. Potomac, Maryland: Lawrence Erlbaum Associates.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

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

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Simon, H. A., & Lea, G. (1974). Problem solving and rule induction: A unified view. In L. W. Gregg (Ed.), *Knowledge and cognition*. Potomac, Maryland: Lawrence Erlbaum Associates.

Wünscher, T., & Ehmke, T. (2002). *Drehmomente sehen. Eine Lerneinheit mit interaktiven Geometrie-Modulen für den Physikunterricht*. Universität Kiel: Leibniz-Institut für die Pädagogik der Naturwissenschaften (IPN). URL: http://www.ipn.uni-kiel.de/abt_physik/drehmomente [15.06.2008].