Exploring a Domain with a Computer Simulation: Traversing Variable and Relation Space with the Help of a Hypothesis Scratchpad.  

Wouter van Joolema* and Ton de Jonge

*Department of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
*Department of Philosophy and Social Sciences, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands

Abstract: Computer simulations provide a challenging opportunity to create learning environments in which learners are free to explore the domain and discover the domain properties themselves. Contemporary theories of learning state that knowledge acquired and constructed within such an exploratory environment will be rooted more deeply within the learner’s knowledge structures. On the other hand, it has also become clear that offering a simulation to the learner without offering any additional support may result in learners getting “lost” in the simulation environment, not learning very much. Therefore, additional support is deemed necessary for simulation learning environments.

The current chapter describes a software instrument (a “hypothesis scratchpad”) that can be offered to the learner in order to support the process of hypothesis formation. The design of this instrument is based on an analysis of hypothesis space, one of the two search spaces from the theory of Klahr and Dunbar [14], who describe discovery as search in two related spaces, a hypothesis space and an experiment space. A study was performed in which the hypothesis scratchpad was used to influence the learners search processes in variable space and relation space, the two subspaces of hypothesis space. The moment of stating hypotheses (before doing experiments or during a series of experiments) and the guidance for variable space search (starting with instances or directly going to general variables) was varied in this study. It appeared that the learners who were prompted to state hypotheses before doing experiments stated more hypotheses and were better judges of the truth value of these hypotheses than students who stated hypotheses while doing experiments with the simulation. Learners who were searching variable space only at the level of general variables, chose relations at a more precise level than learners searching variable space, starting with instances before going to the general concepts.

Keywords: Computer Simulations, Exploratory learning, Intelligent Tutoring Systems

1Part of this paper is based on research conducted in the project SMSLE. SMSLE is a R&D project partially funded by the CEC under contract D2607 within the main phase of the DELTA programme.
Introduction

Nowadays, the recognition grows that learning should not only take place in "traditional" learning settings in which a teacher exposes information to be learned to the learner, but also in more "exploratory" settings in which learners are allowed to interact with an object of study and discover the properties of the domain themselves. This "constructionist" approach would result in knowledge, rooted deeper within the knowledge structures of the learner [6]. Computer simulations are an excellent means to create such exploratory learning environments [1,7]. In simulation environments the learner can interact with the simulation by changing variables and observe the effects of these changes.

The basic task for a learner in such an exploratory environment is to discover the rules that underly the behaviour of the simulation. In most cases the rules that are to be discovered by the learner are not the exact quantitative, programmed rules that drive the simulation, but rather some rules representing a more abstract, often qualitative, understanding of the domain [13]. In other words, the goal of learning with an exploratory simulation environment is to build a quantitative or qualitative conceptual model of the domain. In order to construct such a model, a learner must be able to state hypotheses, design experiments and draw conclusions from these experiments [21]. These study processes have proven to be problematic for many students [8,20].

Exploring a simulation as dual space search

Constructing a conceptual model capable of describing and predicting the essential features of a simulation, based on the simulation results, is a rule-discovery task. This means that a learner has to discover rules (or relations in our terminology), describing the behaviour of variables as functions of each other. These relations can be used for predicting the simulation's behaviour in new situations. Klahr and Dunbar [14] (see also [5,15,25]) have characterised this rule-finding task as a search task in two distinct, but related, search spaces: a hypothesis space and an experiment space.

Hypothesis space is a search space consisting of all rules possibly describing the phenomena that can be observed in the domain. Experiment space is the space consisting of experiments that can be performed with the domain and the outcomes of these experiments. The search processes in these two search spaces are considered to be more or less similar to searching a problem space as part of a problem solving process in the sense of Newell and Simon [19].

Two exploratory strategies: theorists and experimenters

Greeno and Simon [5] distinguish two types of strategies for traversing the two search spaces: a bottom-up strategy, starting with an analysis of the data, found in experiment space, with the purpose to generalise the data to a rule and a top-down strategy, starting with choosing a node in hypothesis space and testing this hypothesis with data found in the instance space. Klahr and Dunbar [14] use a similar distinction and label these strategies as "experimenter" and "theorist" strategies respectively.

---

2 Greeno and Simon [5] use the terms rule space and instance space respectively. The terms in the text have been introduced by Klahr and Dunbar and are especially suitable for describing discovery tasks related to simulations or devices. Greeno and Simon's terms can have a more general use e.g. [15].
The two discovery strategies were observed experimentally by Klahr and Dunbar [14]. They found that subjects who were not able to state the correct hypothesis before any experiment was carried out became "experimenters" and students able to state the goal hypothesis were "theorists". Moreover, they found that in 56% of the cases students failed to draw the right conclusions from disconfirming experiments: i.e. hypotheses were retained incorrectly on the basis of a negative experimental result. A reverse effect, learners rejecting hypothesis without a disconfirming experiment was also observed. A possible explanation for this behaviour is that the relation between hypothesis space and experiment space was not clear to the learners.

It follows from the above that, in order to understand discovery learning in a specific domain, one needs to investigate the properties of the associated hypothesis space and experiment space and the relation between the two. This relation sometimes is trivial, as in the case of the task described by Qin and Simon [25] where the relation between two variables was to be discovered. In their study, "instance space" consisted of a set of four values for the two variables. A hypothesis could immediately be tested by examining the data, the result was always unambiguous: true or false. Qin and Simon do not report on incorrect use of experimental data.

In Klahr and Dunbar's study [2, 14] the domain was the operation of a programmable device. Hypothesis space consisted of possible rules that described the operation of a certain button, experiment space consisted of all possible programs. An important difference with the Qin and Simon study was that not all experiments could render an unambiguous result. In the current study we add an extra complexity, viz., the domain contains a hierarchical structure of different kinds of variables.

The properties of hypothesis space for simulations

The properties of hypothesis space for simulations are determined by the shape hypotheses about a simulation model take. A hypothesis is a relation between two or more (conceptual) variables [12]. This implies that hypothesis space has a variable dimension and a relation dimension.

Both of these dimensions have a structure, i.e., either dimension is a structured subspace of hypothesis space, hence we will speak of relation and variable space, rather than dimension. In [9, 15] the structures of variable and relation space have been elaborated. Both subspaces have a hierarchical organisation. For relation space going down in the hierarchy means improving on precision of a relation. Precision is, in this case, defined by using the relation as a predictive engine, the less alternatives that are allowed by a relation, the more precise it is. For example, the relation "if A increases B also increases" is less precise than "if A doubles, B also doubles". A similar concept has been defined by Pöltzer and Spada [22, 23] who distinguish three levels of precision: qualitative relational, quantitative relational and quantitative numerical. Going up in the variable hierarchy means improving on generality, meaning improving the scope of validity of a hypothesis, since a hypothesis containing more general variables can be applied to more instances. For example, a hypothesis can be stated about the concept of "energy" in general, but it may also be limited to the less general concept "kinetic energy", or even "the kinetic energy of particle A". At the

Here is a case of a finite and small experiment space where the value "true" may be assigned unambiguously. In general this is not possible, only disconfirmation of hypotheses is possible with certainty [24].
lowest level of generality, variables are instantiated with variables in the simulation model. This "runnable" model defines the properties of experiment space.

Additionally, we may divide hypothesis space in a number of regions:

- the universal hypothesis space, the (theoretical) space of all possible hypotheses that can be constructed.
- the learner hypothesis space, the space of all hypotheses that the learner knows of to exist.
- the learner search space, the space of hypotheses the learner actually considers for describing the domain, that is those hypotheses actually generated by the learner.
- the conceptual model, the goal set of the search, viz., the set of "true" hypotheses.

In Figure 1 an example of a configuration of these spaces have been drawn. It is clear that the learner hypothesis space is necessarily a subset of the universal hypothesis space, and that the learner search space is a subset of the learner hypothesis space. It is not necessarily true that the conceptual model is a subset of the learner search space, or even of the learner hypothesis space. If the conceptual model lies, partially or wholly, outside the learner search space, the learner will have to extend this search space, because the student will not be able to "reach" the hypotheses outside the learner search space. This "stretching" operation on the learner search space will require extra conceptual effort. Therefore, one would expect that offering support on exploration should primarily be directed at the creation and extension of the learner search space and, if needed, the learner hypothesis space.

Research questions

In the study described in this chapter students worked with a computer simulation of a chemical experiment trying to construct a "mental" conceptual model of the model behind the simulation. We aimed to influence the learner search space by providing students with specified (general or instantiated) variables and to have students "articulate" their search space by letting them indicate relations between these specified variables, either before or while they conducted experiments. Entering relations between specified variables (thus
constructing hypotheses) was done with the help of a so-called hypothesis scratchpad. As a dependent variable we looked at the quality of the exploration, expressed in the activity level of the students, the final result and the effectiveness of the search process.

*Influencing the search in relation space*

It was tried to influence the learner's search in the relation space by presenting the *elements* of this dimension in the form of lists with types of relations. For entering relations a *hypothesis scratchpad* was introduced. Using this (software) tool students could enter relations of different domain of applicability and precision between variables, also present on the scratchpad. The combination of variables and relations constitutes the learner's hypothesis space. Half of the students were forced to articulate their "learner search space" before conducting experiments, the other half of the students could enter relations between the variables in the assignment at any moment. By forcing students to articulate their search space before conducting experiments we expected more planned and systematic experimentation behaviour, also resulting in a lower number of experiments conducted [14]. Also, we expected these students to include more precise relations than the students who could construct their search space more opportunistically. In a previous experiment we found that students who stated relations in the course of experimenting with the help of a (mocked-up) scratchpad, mainly chose very global relations [12].
Influencing the search in variable space

For all students in our study, we provided a structure for the search in variable space by introducing assignments that invited students to investigate the relation between two (or sometimes three) specified variables. The subsequent assignments were chosen in such a way that together all important variables from the domain were covered. For half of our subjects these assignment always contained general variables, the other half of the subjects started with variables that had obvious instantiations in experiment space, and ended with variables at the general level. This difference between general and instantiated variables is not present in the domains investigated by Klahr and Dunbar [14] and Qin and Simon [25]. By giving variables at an instantiated level and following this by offering the general variables, we expected to help student in finding the link between instantiated and general variables, i.e., in establishing a better link between experiment and hypothesis space. This should be reflected in a better design of experiments and better conclusions to be drawn from them. During working on an assignment, the only variables present on the scratchpad were the variables occurring in the assignment, so a learner could only state hypotheses about those variables.

The resulting "true" hypothesis space

For each of the hypotheses stated we asked students to indicate whether they thought a hypothesis was "true", "false", or they "don't know". We regarded the lists of hypotheses with these values attached to them as representing the current knowledge of the learner of the domain. Therefore, we could see these lists at the end of the session as the learning result.

Method

4SEE, an integrated simulation environment for exploring error analysis

The experiment was conducted with a learning environment called 4SEE (Statistics Simulation System as a Supportive Exploratory Environment), developed at the Eindhoven University of Technology. This environment consists of a simulation of a chemical experiment and a hypothesis scratchpad which the learners can use to note down hypotheses. In literature, learner instruments for support on stating hypotheses have been described earlier [17,26,27]. The version of 4SEE (version 2) used in the current study is a rewritten version of 4SEE version 1, used in our previous study [12]. The simulation model itself has not changed, but the learner interface and the hypothesis scratchpad have undergone radical changes. The main change is that the hypothesis scratchpad has been integrated in the learning environment.

The simulation

The simulation used in the experiment simulates a titration experiment where the emphasis is on analyzing measuring errors made in chemical (and other) experiments. In this domain the relation between the measuring errors made in (partial) measurements within a measuring process and the total error at the end of such a process is of primary importance. This law of error propagation indicates ways to decrease the error in the final result of the chemical experiment. 4SEE is designed to perform simulated titration experiments and observe the relation between measuring errors occurring in this experiment (which are also simu-
lated). The goal is to discover the properties of the relations between the different kinds of measuring error and other experimental parameters. The simulation allows for repeated experiments and keeps an inspectable history of the experiments performed.

The hypothesis scratchpad
The hypothesis scratchpad used in this study consists of three lists, containing ingredients for a hypothesis: one list contains variables, one contains relations and one contains conditions. In a separate window the complete hypothesis under development is displayed. A student may enter a hypothesis by selecting a relation and then filling in its empty slots by selecting variables from the variable list or typing in values. Optionally, the student can limit the domain of applicability of the relation by adding a condition, which may itself contain slots to be filled in by values or variables. The result of the operations sketched here is a complete sentence stating that a relation holds between variables (see Figure 2).

Once completed, a student can save the hypothesis and create another one, or proceed with the simulation. Each hypothesis saved is entered on a list which the student can inspect. Once a hypothesis has been added to the list, it can not be deleted or changed any more. Student can mark their progress on the list in two ways: a hypothesis on the list can be marked to be "under investigation" and, moreover, it can be marked as "true", "false" or "don't know". The hypothesis list can always be inspected, and the marks on the list can be changed at any moment.

Experimental set-up

The experiment followed a 2x2 factorial design with assignment type and enforced experimental strategy as independent variables.

In the "assignment type" manipulation one group of subjects received assignments containing variables at the instantiation level, followed by an assignment with more general variables. Another group received only the general level variables. The subjects in the "instance" group received fifteen assignments, the "general" group received four assignments. An example of a variable at instantiation level is "the measuring error of the balance"; an example of a general variable is: "the partial error, participating in a calculation of another error."

For manipulating the "strategy" variable two different versions of the program were used. One group of subjects, which will be labelled the "experimenter" group, received the simulation with the possibility to switch to the hypothesis scratchpad at any moment to create new hypotheses. Another group, called the "theorist" group was forced to start a new assignment with constructing a hypothesis list containing all hypotheses that the subject thought to be likely candidates for correct hypotheses. After subjects had completed this hypothesis list they could start an experiment and switch back and forth between the hypothesis list and the simulation. There was a possibility of adding new hypotheses to the list, but the subjects in the theorist condition were instructed to use this possibility only as an emergency escape, in case all hypotheses on the scratchpads proved to be false.

*The terms "experimenter" and "theorist" are the same as used by Klahr and Dunbar (1988). Here they are used a little differently: Klahr and Dunbar use the terms for behaviour that occurs spontaneously, here they are used to indicate behaviour imposed by the learning environment, at least for the "theorist" group. Subjects in the "experimenter" group are of course free to behave like theorists.
Subjects

Subjects were 37 first year students of chemistry at Eindhoven University of Technology. They had had an introductory, formal course in measuring analysis as part of their normal study program. They were randomly selected by their teachers to participate in the study, which took place at a point in time where the simulation would normally be used as part of a chemistry course. Groups were assigned randomly to conditions. For their participation, the subjects were excused from a lab assignment they normally would do during the time they participated in the study. For an overview of the experimental design and the number of subjects in each condition see Table 1.

Procedure

The subjects received a written instruction on the use and purpose of the computer simulation. In this instruction it was emphasized that the goal of the simulation was to discover relations between variables and to enter these, and any alternatives considered, on the hypothesis list. After the subjects had time to read the instruction a demonstration of the use of the simulation was given. Furthermore, the subjects received a set of assignments, each of the form: "investigate the relation between:" followed by the names of two or (in some cases) three variables. The variables in the assignment were the only ones present on the scratchpad. The scratchpad contained a button to switch to the next assignment, in which case a new set of variables (belonging to the next assignment) was loaded into the scratchpad. Subjects themselves determined when an assignment was completed. The final, annotated, hypothesis list on the scratchpad was considered to be the result of an assignment.

The duration of a session was three hours. In this time subjects worked with the simulation: they performed experiments with it and stated hypotheses. The subjects were informed that there was no need to complete all assignments, they completed as many assignments as time allowed.

Data collection

The data collected consisted of the log files recorded during the experimental session. Every action by the learner was recorded, yielding a complete overview of the experiments performed and hypotheses stated. Each hypothesis was assessed on correctness, precision and domain of applicability. Part of this assessment was done automatically, using a computer program QMaPS, a general reasoning system, using a conceptual model of error analysis. QMaPS is discussed in more detail in [10].

All subjects were given a post test, consisting of 6 multiple choice items in which subjects had to apply knowledge from measurement analysis, which they could have acquired from working with the simulation. Also, the test contained two questions in which subjects were asked to mention as many relations as they could think of, prompted by the question: "mention as many relations between the variables .. and .. you think a beginner on this subject might consider as hypothesis when exploring this simulation". The slots in this question were filled in with (general) variables given in the assignments for the exploration of the simulation.
Table 1 Number of subjects in the different experimental conditions.

<table>
<thead>
<tr>
<th>strategy</th>
<th>assignment type</th>
<th>instance</th>
<th>general</th>
</tr>
</thead>
<tbody>
<tr>
<td>experimenter</td>
<td></td>
<td>$N = 10$</td>
<td>$N = 9$</td>
</tr>
<tr>
<td>theorist</td>
<td></td>
<td>$N = 9$</td>
<td>$N = 9$</td>
</tr>
</tbody>
</table>

Results

The data collected were analyzed on the following aspects: level of activity of the subjects, the precision of the hypotheses stated (which is a property of the relations chosen), the quality of the exploratory activities of the subjects and the results of the exploration, expressed by the final list of hypotheses on the scratchpad at the end of each assignment. An analysis of the level of generality of hypotheses stated by the subjects (a quality of the variables in the hypothesis), was not possible due to the fact that this level was enforced by the assignment type manipulation. A qualitative analysis showed that, in general, subjects in the experimenter group indeed behaved as experimenters (see Note 4). Finally also the result of the post test was analyzed.

Activity level

The activity level of the subjects was measured using three indicators: the number of hypotheses stated, the number of experiments performed with the simulation and the number of assignments completed by the subjects. In Table 2 these variables are depicted for the various groups. For the total number of hypotheses stated, a comparison between the instance group and the general group is not possible, because of the larger number of assignments given to the instance group. Therefore, the analysis was performed separately for both groups.

No significant differences between groups in the number of experiments performed. Two non-significant trends are found: subjects in the theorist group perform ($p = 0.16$) fewer experiments than subjects in the experimenter group and "general" subjects perform less experiments than "instance" subjects ($p = 0.08$). Subjects in the theorist group produced a higher number of hypotheses than the experimenter group in the instance condition. For the general condition a similar, not significant ($p = 0.073$), trend was found.

The number of hypotheses per assignment is also greater for theorists than for experimenters, as depicted in Table 3. There was also a main effect of the assignment type: the general group stated more hypotheses per assignment than the instance group.

Another indicator for the activity level is the number of assignments completed. For this number only a comparison is possible within the strategy manipulation, for the instance condition, as a consequence of the smaller number of assignments given to the general group. No significant differences were found, on average subjects completed about 3.5 assignments in both the theorist and experimenter condition.
Table 2. Activity level of subjects, expressed in the number of hypotheses stated and the number of experiments performed.

<table>
<thead>
<tr>
<th>strategy</th>
<th>average number of hypotheses per subject</th>
<th>average number of experiments per subject</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>assignment type</td>
<td>assignment type</td>
</tr>
<tr>
<td></td>
<td>instance</td>
<td>general</td>
</tr>
<tr>
<td>experimenter</td>
<td>8.4</td>
<td>8.3</td>
</tr>
<tr>
<td>theorist</td>
<td>19.6</td>
<td>13.2</td>
</tr>
<tr>
<td>total</td>
<td>14.3</td>
<td>10.8</td>
</tr>
</tbody>
</table>

Treatment effects of strategy manipulation:

- Instance group:
  - F(1,17) = 7.27, p < 0.05
- General group:
  - F(1,18) = 3.62, n.s.

Choosing relations: precision of hypotheses stated

The precision of the hypotheses entered on the scratchpad was measured on a discrete scale from zero to two. Level 0 was a very imprecise level ("there is a relation"), level 1 corresponds roughly to the qualitative relational level, level 2 to the quantitative relational level, as defined by Ploetzner and Spada [23].

In assessing the precision on which hypotheses are stated, the maximum precision per assignment was analyzed, rather than the average precision. Two variables were observed: the maximum precision of all hypotheses stated by the subject and the maximum precision of the hypotheses marked as true, being those hypotheses which were eventually used by the learner for modelling the relation between the variables defining the assignment, independent of the actual truth value of the hypothesis in the conceptual model. The results of this analysis are summarized in Table 4. A significant effect of the assignment type on both types of maximum precision is found: the "general" subjects use more precise hypotheses than the instance subjects. Table 4A also shows a small significant effect of the strategy manipulation: theorists state more precise hypotheses than experimenters. However, in Table 4B this effect has disappeared. This means that the theorists reject more of the (higher number of) hypotheses that they stated at the more precise levels, when compared to experimenters.

Quality of exploration

As an indicator for the quality of exploration we used the relevancy of an experiment for the assignment and the indication by the student if s/he was testing hypotheses.

There was a rather low score on the direct relevance of experiments, on average 22% of the experiments were of direct relevance. There is no effect of either condition on this aspect. A conclusion that students are poor experimenters would be too preliminary: there should be room for free exploration. Furthermore, the score used was very sensitive and did not discriminate between moderate and severe deviations from an ideal behaviour.
Table 3 Number of hypotheses entered per assignment

<table>
<thead>
<tr>
<th>strategy</th>
<th>assignment type</th>
<th>instance</th>
<th>general</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>experimenter</td>
<td></td>
<td>1.78</td>
<td>3.16</td>
<td>2.18</td>
</tr>
<tr>
<td>theorist</td>
<td></td>
<td>2.61</td>
<td>3.86</td>
<td>2.94</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>2.28</td>
<td>3.58</td>
<td>2.64</td>
</tr>
</tbody>
</table>

Main effects:
- strategy:
  - $F(1,183) = 13.54, p < 0.01$
- assignment type:
  - $F(1,183) = 30.04, p < 0.01$

Another indication of the quality of exploration is the students' own indication of testing hypotheses. We scored for every student the fraction of experiments that was designed to test one or more hypotheses, according to the markings on the hypothesis list. In Table 5 the results of an analysis of variance for this score is presented. A positive effect of the experimenter/theorist condition is found: on average, students in the theorist condition devote a larger percentage of their experiments to testing at least one hypothesis than experimenter students. No significant effect was found from the instance/general condition.

Results of exploration

The results of the exploration process can be expressed in the number of true hypotheses that have been found and marked as true by the students, i.e. the number of correctly accepted hypotheses. Of course, also the other hypotheses, viz., the hypotheses that were rejected, correctly or incorrectly, and the hypotheses that are not true and have been accepted, determine the success of the exploratory session with the computer simulation. For the purpose of this analysis all hypotheses stated by the subjects are classified according to their truth value in the (expert) conceptual model and the truth value assigned by the learner, yielding values from the set "correctly accepted", "correctly rejected", "incorrectly accepted" and "incorrectly rejected" (and categories for the case that the learner stated he didn't know the truth value). There was a significant effect of the theorist/experimenter condition on these scores. ($\chi^2(5, N=438) = 33.1, p < 0.01$, for a contingency table of score vs. theorist/experimenter). It appeared that for 60% of the hypotheses stated by theorists, the value assigned by the learner matched that of the conceptual model, opposed to 44.9% for experimenters. Experimenters accepted more false hypotheses than theorists (42% vs. 20.5%). No such effects were observed for the instance/general condition.

An analysis of variance, for the total number of hypotheses correctly accepted per student revealed that, on average, students in the theorist group stated and accepted 4.5 correct hypotheses, whereas students in the experimenter group stated and accepted 1.7 correct hypotheses ($F(1,35) = 6.3, p < 0.05$). No such main effect was observed for the instance/general condition.
Table 4 Maximum precision per assignment, both for all hypotheses stated and for the hypotheses marked as true.

<table>
<thead>
<tr>
<th>A</th>
<th>maximum precision of hypotheses stated for each assignment, averaged per subject over assignments.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>assignment type</td>
</tr>
<tr>
<td></td>
<td>strategy</td>
</tr>
<tr>
<td>experimenter</td>
<td>0.96</td>
</tr>
<tr>
<td>theorist</td>
<td>0.99</td>
</tr>
<tr>
<td>total</td>
<td>0.97</td>
</tr>
</tbody>
</table>

main effects:
- strategy: $f(1,35) = 5.03, p < 0.05$
- assignment type: $f(1,35) = 7.66, p < 0.01$
- interaction: $f(1,35) = 4.30, p < 0.05$

<table>
<thead>
<tr>
<th>B</th>
<th>maximum precision of hypotheses marked as “true” for each assignment, averaged per subject over assignments.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>assignment type</td>
</tr>
<tr>
<td></td>
<td>instance</td>
</tr>
<tr>
<td>E</td>
<td>0.80</td>
</tr>
<tr>
<td>T</td>
<td>0.52</td>
</tr>
<tr>
<td>total</td>
<td>0.66</td>
</tr>
</tbody>
</table>

main effects:
- strategy: $f(1,35) = 0.047, n.s.$
- assignment type: $f(1,35) = 5.40, p < 0.05$
- no significant interaction effect.

Post test

The results of the post test are presented in Table 6. The results of the 6 multiple choice questions and of the open questions are presented separately. For the open questions the number of relations mentioned is listed. An analysis of variance revealed no significant differences for any of these variables or conditions.

Of course just listing the number of relations mentioned does not reveal the complete results of the open questions; also the nature of the relations mentioned should be taken into account. These results are anecdotal.

It appeared that the precision level relations the students mentioned for these two questions was almost invariably 1, i.e., at the qualitative relational level. In only three cases students mentioned more precise relations. Furthermore, some (5) students misinterpreted the questions and mentioned instances of the variables given in the assignment instead of relations. One student, in the experimenter/general condition, gave not only a set of relations to investigate but a complete experimental plan to determine the correct relation.

Conclusions

The first research hypothesis stated that a search of hypothesis space before the start of an experiment would result in a larger learner search space, which would result in both more precise results of the exploration and in more efficient experimentation behaviour. The first part of this research hypothesis was confirmed: the average number of hypotheses stated per assignment was greater for theorists than for experimenters, meaning that the students in the theorist group used a larger search space. Also, the maximum precision of the hypotheses stated in the learner search space was significantly higher for theorists than for
Table 5 Average fraction of experiments performed per subject, designed to test one or more hypotheses.

<table>
<thead>
<tr>
<th>strategy</th>
<th>assignment type</th>
<th>instance</th>
<th>general</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>experimenter</td>
<td>0.22</td>
<td>0.43</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td>theorist</td>
<td>0.47</td>
<td>0.68</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>0.36</td>
<td>0.84</td>
<td>0.44</td>
<td></td>
</tr>
</tbody>
</table>

main effects:
- strategy
  \[ F(1, 35) = 4.45, p < 0.05 \]
- assignment type
  \[ F(1, 35) = 3.02, n.s. \]

experimenters. However, when only the hypotheses that were marked by the students as being "true" were considered, this effect disappeared. This shows that our assumption that a greater learner search space automatically leads to a better, more precise, conceptual model of the domain, constructed by the learner is not correct, the learners will also need support on searching their learner search spaces and selecting hypotheses from them. The theorists, however, did find more correct hypotheses and accepted more of them as true, than experimenters. Therefore the conceptual models constructed by the theorists were not better in precision than those of the experimenters, but they contained more correct relations.

Concerning the second inference from the research hypotheses, regarding the number of experiments performed by the theorists, we found a non-significant trend in the expected direction. We did find that, per student, a greater percentage of experiments performed by theorists was, according to their own sayings and when compared to the experimenters, performed to put one or more hypotheses to the test.

The students in the theorist condition proved better judges of the truth value of a hypothesis. In significantly more cases they correctly rejected a false hypotheses or accepted a true one, than the student in the experimenter group. The theorists also rejected more hypotheses in general, which is extra support for the assumption that the theorists explored a wider range of hypotheses.

Overall, concerning the theorists/experimenter condition, the conclusion that the stimulus for generating hypotheses before experimentation has a positive effect on the exploratory behaviour seems to be justified. On the average, the theorists build a larger learner search space, which also contains more precise hypotheses and the final result of exploration is also better in terms of the number of correctly accepted hypotheses. A problematic point is retaining the level of precision of hypotheses, from the initial search in hypothesis space to the final set that the students mark as being true.

The better exploratory behaviour and the resulting larger set of true hypotheses found by theorists, however, was not reflected in better results on the post test, which showed no differences between the different experimental groups. Probably, this is a consequence of the fact that the experiment took place in a short period of time: it may be a little optimistic to expect significant differences in learning results in three hours. An explanation can be sought in the precision level of the conceptual models constructed by the students, which did not differ for theorists and experimenters. For scoring better on the post test, the
students needed more precise relations than they were able to find in the time given.

An improvement, however, on the process of discovery, will inevitably have its effects on the longer term. This is supported by research indicating that generic discovery behaviour and scientific reasoning skills are of primary importance for discovery learning [16].

As an absolute figure, the average maximum precision of the hypotheses the students use to model the domain is rather low. This also occurred in the post test where students mainly mentioned relatively imprecise hypotheses that should be considered for investigation. This low precision accounts for the absence of effects of the better behaviour showed by the theorist learners on the results of the post-test multiple-choice questions, since answering these questions requires knowledge on a higher level of precision.

The low precision level observed can be explained by assuming a, what we like to call, "fear of rejection": students choose those hypotheses of which they expect that they will not be rejected. This would be a counterpart of the "confirmation bias" observed in some other studies [3,4,18,28], where students often chose experiments which could not disconfirm their hypothesis. In this case, students chose hypotheses of which they know that there will not be many experiments to disconfirm them. The confirmation bias thus has its impact on experiment space search, the fear of rejection on hypothesis space search.

The second research hypothesis stated that traversing variable space bottom-up from the instances to the general variables would result in a better articulated relation between hypothesis and experiment space. This should be expressed in a better design of experiments and in drawing better conclusions from these experiments. Such effects could not be found in the experiment described in the current chapter.

Surprisingly, an effect was found on the precision of hypotheses stated as a function of the instance/general condition: students in the general condition, stated hypotheses of a higher precision than students in the instance group. This occurred both for all hypotheses stated and for the hypotheses actually used for modelling the simulation, the hypotheses marked as "true". Apparently the students in the general group were more optimistic about the validity of their, precise, hypotheses. This would seem to be in contradiction with the "fear of rejection" hypothesis, expressed above, since precise hypotheses about general variables have a high chance of being rejected. A way to explain this apparent contradiction lies in the assumption that the relation between hypothesis space and experiment space is
or students who only use the general variables. In this case it would not be clear
thesis is likely to being rejected, since the consequences of experiments are not
the relation between the two search spaces is not. Of course, this remains
further research, especially because the current study was not able to reveal a
in the relation between hypothesis and experiment space for the instance and the
dition.

e case interactions were found between the theorist/experimenter condition and
el/general condition. Also conditions had effects on different aspects of the search
investigated. This means that the influences of both instructional measures, as
in the current experiment are relatively independent of each other. This supports
ical assumption behind the decomposition of hypothesis space in independent
relation spaces. The conclusion seems to be justified that the prior selection of
arily influences the, domain independent, discovery skills, like stating and
oposes, whereas the instance to general measure has its impact on the more
pecific aspects of the simulation. It must be remarked that the latter influence is
than the first.
ous study [12], we have emphasized that the hypothesis scratchpad had a dual
iding support for the learner and providing the learning environment with infor-
out the current knowledge state of the learner. In the current study we introduced
alysis of hypotheses and experiments by the system QMaPS [10]. It turned out
ystem can be used to assess the precision and generality of hypotheses. Further
eed to develop rules for assessing the quality of experiments, such as the fact if
ent is readily suited for testing a hypothesis.
ortant result of the current study is that the exploratory behaviour of the students
fluenced by the design of the learning environment: the "theorist" group behaved
theorists in the sense meant by Klahr and Dunbar [14]. This leads to the con-
t that there are challenging opportunities to influence or, better, improve the student's
cesses in hypothesis space. Special attention should be given to the relation
thesis and experiment space. Moreover, in order to be able to stimulate stu-
covering the simulation properties more in depth, i.e., by finding more precise
of the motives behind making hypothesis space moves is needed [11].

dgments

e due to Gert-Jan van Rootselaar for programming the simulation and to Dr. J.C.
nd Dr. M.C.A. Donkersloot, for their cooperation with this research and their
is for improving the simulation.

18

es, NY: Prentice-Hall.

de, New Jersey: Lawrence Erlbaum associates.


