Pergamon

Studies in Educational Evaluation, Vol. 22, No. 4, pp. 341–362, 1996

0191–491X/96 $15.00 + 0.00

S0191-491X(96)00019-3

MEASURING INTUITIVE KNOWLEDGE IN SCIENCE: THE DEVELOPMENT OF THE WHAT-IF TEST

Janine Swaak and Ton de Jong[1]

Faculty of Educational Science and Technology,
University of Enschede, The Netherlands

Introduction

Learning in discovery environments, such as simulations, is considered to be different from the acquisition of knowledge in a more expository instructional context. In the discovery environment knowledge is acquired in a process of active knowledge construction, whereas in expository teaching knowledge is transferred from teacher to student. Discovery learning with simulations is, therefore, supposed to lead to acquisition of knowledge that is qualitatively different from knowledge that is acquired in more traditional instructional situations where an emphasis on the formal aspects of domains is present. Currently, however, there appears to be a mismatch between the context of learning and the resulting qualities of knowledge, and the methods used to assess these qualities of knowledge.

Thomas and Hooper (1991), for example, classified and analysed simulation studies according to the instructional function for which the simulation was used. The conclusion most relevant for the present work is that, in the authors' words, "the effects of simulations are not revealed by tests of knowledge (....)" (p. 479). The knowledge tests used in most of the simulation studies reviewed consisted of items involving "definitions", "recognition", "recall", and "direct application" (pp. 500-501). Thomas and Hooper continue that effects of simulation can instead be detected by "tests of transfer and application" (p. 479). They further conclude that simulations can best serve as "experiencing programs" which give students the opportunity "to gain an intuitive understanding of the learning goal" (p. 499). Thomas and Hooper do, however, not indicate what they mean by "tests of knowledge", why and how they think an intuitive understanding is gained, and how intuitive understanding can be assessed.

As another example, Shute and Glaser (1990) compared a test with multiple choice items and a test with short answer items in a large scale evaluation of an intelligent simulation on Economics, called Smithtown. They found that the multiple choice questions did not differentiate students who had worked with the simulation from students who had followed lectures on the same domain. The short answer questions yielded a significant difference between the two experimental groups. Shute and Glaser (1990) do not try to explain these results and do not specify the differences in contents between the two used item formats. The results, however, indicate that, whatever the reason, the multiple choice items do not tap the same aspects of knowledge as the short answer items, and that, therefore, testing format is an important factor in determining the success of instructional practice.

In sum, in existing studies we find indications for specific relations between the nature of learning, the qualities of the resulting knowledge, and the methods to measure the acquired knowledge. The objective of our work is to have a closer look at the relations between the nature of simulation-based discovery environments, the learning elicited, the knowledge resulting from discovery learning, and the methods to assess the resulting knowledge.

Learning in Simulation Based Environments

In this paper we focus on simulation based learning environments. The primary mode of learning in simulation based learning environments is (scientific) discovery learning. Simulation environments are environments in which learners are invited to discover the rules and laws of a domain. This means that in principle the domain is not offered directly to the learner, but the learner has to *infer* characteristics from the domain, from the information offered. The environments we present in the current paper are developed with the SMISLE[2] system. These environments carry a number of specific features (de Jong, van Joolingen, Pieters, van der Hulst, & van der Hoog, 1992; de Jong & van Joolingen, 1995). At the heart of such an environment is always the simulation in which the subject matter is simulated. In this simulation, learners can manipulate variables and see the consequences of their manipulations in dynamic outputs. Furthermore, the simulation environment is enriched with support measures added to the simulation. The support measures that can be created with SMISLE include explanations, assignments and model progression. The explanations consist of explanations of variables of the domain. Students can at own will, choose to look up an explanation in which case a description of the variable is given. Assignments are small exercises that aim to help students structure their learning and to direct them to important phenomena in the domain. Among these are investigation assignments which ask the student to investigate the relation between two variables of the domain by manipulating the variables in the simulation, and prediction assignments in which the student is asked to predict the relation between two variables in the domain by choosing the predicted value of a variable in the simulation. Model progression means that the domain is offered in small subsequent steps, with every step adding new variables to the model. Also, model progression is offered for structuring the learning process (see White & Frederiksen, 1990). Figure 1 shows how the simulation and the support tools might be arranged. The figure shows a screendump of a simulation

environment on the physics topic of collisions (for a full description of this environment see de Jong, Martin, Zamarro, Esquembre, Swaak, & van Joolingen, 1995).

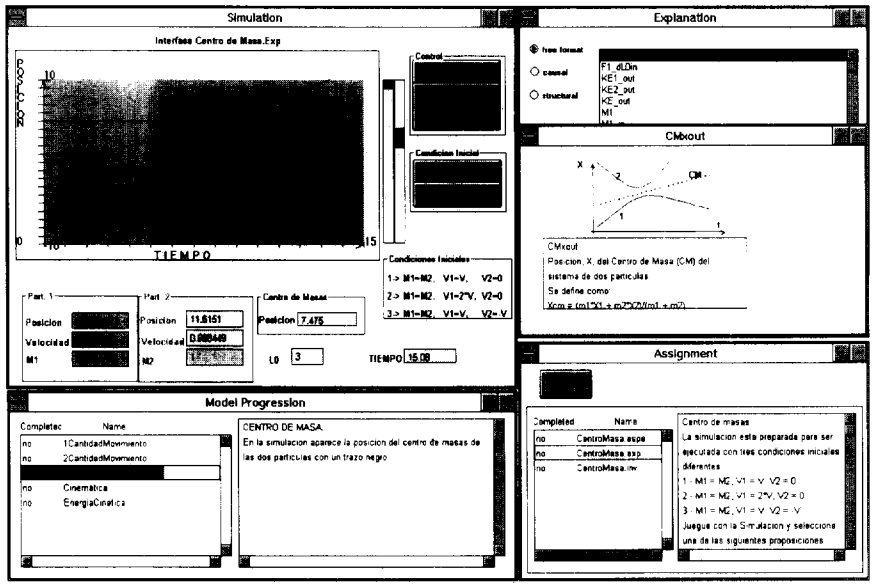

Figure 1: Prototype Interface of a SMISLE Type of Simulated Based Discovery Environment (original Spanish version)

So far, we highlighted very specific features of the SMISLE environments. More general characteristics of simulation based environments are: (a) the richness of the environment, (b) the low transparency of the environment, and (c) the active interaction with the environment.

Simulation based learning environments can be described as rich environments. By saying this we mean that (a) a great deal of information on a domain can be extracted by the learner, that (b) this information can be obtained in several ways, and that (c) the information is usually displayed in more than one representation; a dynamic, graphic representation of the output is generally present alongside animations and numerical outputs (e.g., Chmiel & Tattersall, 1991; Rieber, Smith, Al-Ghafry, Strickland, Chu, & Spahi, 1995). In the simulation, the central part of the learning environment, students can manipulate variables as they like and observe the consequences of their manipulations in dynamic outputs. The learning environment can even be enriched by adding instructional components to the simulation, as is done in SMISLE based environments, but see also e.g., Smithtown (Shute & Glaser, 1990), and Voltaville (Glaser, Raghavan, & Schauble, 1988). These instructional measures are meant to guide the learner through the environment. They usually, however, also make the environment more complex and add methods of information extraction to the learning environment (change component b.), but more information (component a.) or more representations of output (component c.) can also be the consequences of supplementing learning environments with instructional measures.

The second feature of simulation-based environments concerns the relatively low transparency of the learning environment (as compared to text books or hypertexts, etc.).

The less transparent the discovery environment, the less of a "direct view" on the variables and relations is given to the learners (de Jong & van Joolingen, 1996). Consequently, the more is to be inferred or extrapolated. Simulation environments generally have a low transparency; only by adding instructional measures such as explanations, the transparency can be enlarged a little.

The third characteristic of the learning context of this work involves the interaction aspect. The learning session entails an interaction with a simulation based environment. Learners are not supposed to passively absorb information on the domain from the computer screen, but are expected to perform several kinds of actions, in addition to reading and thinking, to make up their own meaningful learning session. Thus, the learning with the discovery environment may accurately be described as an active experience.

The conception of learning with simulation based environments as an experience is very much in line with Norman (1993) who reasons that some environments lead one more toward an "experiential" mode of learning while in other contexts learning can be characterised as "reflective". We argue that the learning environments referred to in this work have some of the important features that foster an experiential way of learning. The reasons for expecting this experiential mode of learning can to a large extent be traced back to the features of simulation based learning environment: the rich, low transparent simulation environment invites the learner to pursue a more action-driven experiential way of learning in which induction plays an important role. Initial confirmation can be found in the work of Chmiel and Tattersall (1991) - as cited in Myers and Davids (1993) - who found that animations and graphical output stimulated an experiential learning mode.

To recapitulate, we argue that the characteristics of the learning context have a considerable influence on the learning processes displayed. For simulation based discovery learning the nature of learning can be described, to a certain extent, as experiential. The features of the learning processes, in turn, will have an impact on the learning products. In the present study we will present a way to assess knowledge acquired during simulation based learning. At this point, we suggest that the knowledge aimed at could be described as *intuitive* knowledge. In the following section, we turn to a selective description of studies on intuitive knowledge, and we will take a look at the characteristics ascribed to intuitive knowledge.

Intuitive Knowledge

Apart from being of a certain type (e.g., operational or conceptual), knowledge can be described by its qualities (de Jong & Ferguson-Hessler, 1996). The quality of knowledge as it is assumed to be gained in simulation based discovery learning is best captured under the term *intuitive*. Fensham and Marton (1991), as referred to in Lindström, Marton, Ottosson, and Laurillard (1993) define intuition as follows: "formulating or solving a problem through a sudden illumination based on global perception of a phenomenon" (p. 264), adding that "it originates from widely varied experience of that phenomenon over a long time" (p. 265). For revealing the intuitions of their subjects, Fensham and Marton used interviewing techniques.

Broadbent, Fitzgerald, and Broadbent (1986) examined knowledge acquired in the control of complex systems and used the term implicit knowledge. In their description of

implicit knowledge, Berry and Broadbent (1988) have related implicit knowledge to implicit learning. They have speculated that "if learners follow an implicit learning mode they probably build up in parallel large number of contingencies rather than following through and evaluating one or more hypotheses" (p. 271). They further argue that the contingencies built up by implicit learning could possibly be represented as a list of procedures, or as Broadbent et al. (1986) have suggested, as form of look-up table. Berry and Broadbent (1988) conjectured that

> in either case a particular set of circumstances or conditions would give rise to a particular response. Whatever the precise nature of the representation, it is likely that some general overall 'pattern-matching' process plays a critical role. (p. 271).

Broadbent and colleagues studied implicit knowledge by looking at the performance of students controlling complex systems. They call their performance tasks cognitive tasks and contrast the performance on these cognitive tasks with verbal knowledge acquired during training with these tasks. The knowledge that could be verbalised was assessed by means of questionnaires consisting of several multiple choice questions or a combination of open and multiple choice questions. Across a series of experiments Berry and Broadbent (1984) found negative correlations between task performance and questionnaire scores. Berry and Broadbent (1988) found no correlations between task performance and written questionnaire scores. In assessing *functional knowledge*, Leutner (1993) has used the same paradigm as Broadbent and colleagues. He looked at the performance in controlling a simulated system during 30 minutes and compared this with verbal knowledge tests. As the scores of the performance measure and the verbal test did not agree, it may be assumed that different qualities of knowledge were tapped by the different measures.

A similar type of knowledge as intuitive, implicit, or functional knowledge is *tacit knowledge* as it is used by Wagner and Sternberg (1985; 1986). These authors considered tacit knowledge to be "(1) practical rather than academic, (2) informal rather than formal, and (3) usually not directly taught" (1986, p.52). To assess tacit knowledge Wagner and Sternberg. constructed a measure consisting of twelve work-related situations. With each situation they associated between nine and eleven response items. Subjects were asked to read a given work-related situation and then to rate each response alternative on a 1-7 point scale by either its quality (i.e., a subjective value judgement) or importance.

To conclude, it appears that intuitive knowledge is a phenomenon frequently referred to in research. However, it seems that in comparison to its claimed relevance, only a moderate effort is invested to assess intuitive knowledge. Despite the under-representation of serious efforts to assess intuitive knowledge, research on interacting with complex simulation systems (e.g., Berry & Broadbent, 1988; Broadbent, et al. 1986; Hayes & Broadbent, 1988; Leutner, 1993), complemented with literature on intuitive knowledge (e.g., Fischbein, 1987; Westcott, 1968), can provide us with at least three, more or less stable, notions on intuitive quality.

The first of these is that the intuitive quality of knowledge is only acquired after *using* knowledge in perceptually *rich*, dynamic situations. We infer that if knowledge is used in rich contexts, experiential learning processes are elicited which lead to intuitive

knowledge. This idea is in agreement with Fischbein's (1987) perception on the acquisition of intuitions. He states that they "can never be produced by mere verbal learning" and that they "only can be attained as an effect of direct, experiential involvement of the subject in a practical or mental activity" (p. 95).

A second finding is that intuitive knowledge is difficult to verbalise. An important hypothesis is indeed that in the interaction with a simulation environment learners are invited to follow an implicit learning mode which leads to knowledge that is hard to verbalise. In a similar vein, Fischbein (1987) states that intuitions are implicit, that they are based on complex selection and inference processes, which are, to a large extent, believed to be unconscious to the individual. Likewise, Westcott (1968) writes that it seems that intuition "occurs when an individual reaches a conclusion on the basis of less explicit information than is ordinarily required to reach that conclusion" (p. 97; cited in Fischbein, 1987, p. 50). Here, it is added that if intuitive knowledge based on information that is not explicitly stated and/or if processes involved in intuitive knowledge are to a large extent unconscious, then intuitive knowledge should be hard to verbalise. Indeed, many of the mentioned studies suggest that there is more to knowledge than only the part that can be verbalised (Berry & Broadbent, 1984; 1988; 1990; Broadbent, et al. 1986; Hayes & Broadbent, 1988; Leutner, 1993).

The third finding is that the access in memory of knowledge with an intuitive quality is different from that of knowledge without this quality. We speculate that the differential access exists alongside differences in verbalisation. We hypothesise that the experiential learning mechanisms tune the knowledge and give it an intuitive quality. Even if it is difficult to verbalise, the intuitive quality causes the access to the knowledge in memory to be more efficient. Van Berkum and de Jong (1991) point at examples in the domain of chess (e.g. de Groot, 1965; Chase & Simon, 1973) which suggest that the phenomenon of knowledge tuning is not limited to operational knowledge (see Anderson, 1987), but that it also extends to more conceptual knowledge. In the authors' words, "Chess masters mainly differ from novices in their 'direct perception' of complex, meaningful chess patterns, and much less in their basic problem solving procedures" (van Berkum & de Jong, 1991, p. 313). In an analogous way, Fischbein writes about the self-evidence and immediacy of intuitions. This entails that an intuition "appears subjectively to the individual as directly acceptable" (p. 200). Fischbein regards "visualisation" as the main factor responsible for the effect of immediacy (p. 103). According to him visualisation may or may not be mediated by an external representation. Fischbein, moreover, mentions that "what one cannot imagine visually is difficult to realise mentally" (p. 103). Though visualisation is critical, Fischbein terms an intuition a theory, not just perception. Yet, he calls an intuition "the analog of perception at the symbolic level".

To sum up, low verbalisability, rich situations and quick perception are the three qualities most frequently found in relation to intuitive knowledge. A certain coherence can be observed between these findings. Several questions remain unanswered, however. So far, there is no agreement on the exact nature of the processes involved in the acquisition of the intuitive knowledge. Even more unclear remains the precise representation of intuitive conceptual knowledge. However, most researchers agree that, whatever the exact nature of the processes involved in the acquisition and whatever the precise representation of intuitive conceptual knowledge, the processes involved in the

manifestation of the intuitive quality of knowledge can be described as a *quick perception of meaningful situations.*

Assessment Methods in Science Education

In the area of science education (in the context of discovery learning with computer simulations) a number of studies have used non-traditional methods for assessing knowledge. Two techniques, asking for predictions and using latency data, seem particularly relevant for measuring intuitive knowledge.

Predictions of Situations

White (1984) compared two groups of 16-17 year old science students studying Newtonian motion with a computer simulation. One group worked with a simulation to which games were added, and the other group worked with a plain simulation. White found that the group with the games outperformed the non-games group on a test which involved qualitative implications of physics laws. In her own words, the problems were "verbal forms of situations that could occur in the Newtonian computer microworld" (p. 81). In more recent work White (1993) used predictions of situations in all her tests. This time she found a difference between 11-12 year old students who had used the Thinker Tools simulation environment during their science curriculum, and students who where taught science in the traditional way. The students working with the Thinker Tools environment outperformed the other group of students.

Latencies

The use of speeded questions in formal, academic domains is a rather new endeavour. Two authors who use this technique are Rieber (e.g., 1990; 1991), and Kieras (1993). Kieras applied this idea in the domain of practical electronics. The engineering students in his experiments studied a series of simple circuit types, each of which performed a specific function. The simple circuit types, which can be combined in more complex circuits, are called building blocks. Half of the students studied relevant building blocks, and half of them irrelevant ones, i.e., with respect to the complex target circuits which followed. The target circuits were accompanied by an explanation and a question. The question always presented a perturbing event, and the answers were a choice of a voltage that either increased, decreased, or stayed the same. The basic measures taken were the time to read the explanation of the target circuit, and the latency and answer of each question. Kieras found that, generally, studying the relevant circuits schemas speeded up both performance measures. However, Kieras's work, mainly points to the problems of using latencies in complex domains[3] and gives practical hints on the construction of items. For example, Kieras found that the data he gathered were "noisy" and that it was hard to detect effects of the experimental manipulations. Furthermore, he warned not to use too many words in the items, to prevent, what he called "word salads".

In work by Rieber, Boyce, and Assad (1990) and Rieber (1990) the use of latencies was more successful. In the study by Rieber et al. (1990) students received a computer-

based, four-part introductory course on Newtonian mechanics. Immediately after each part students received either practice questions with feedback on their responses, a structured simulation in which students received increasing level of control over a free-floating object, or no practice. Furthermore, each of these levels of practice was crossed with three levels of visual elaboration in the introductory course. The contents of the course were either supplemented by static graphics, animated graphics or no graphics at all. Students' acquired knowledge was tested by a 19-item multiple choice test (originally 32 items) intended to measure students' ability to apply rules of the domain. Both correctness scores and latencies were collected. With respect to correctness scores both practice groups outperformed the no practice group. No main effect was found on visual elaboration. With regard to response latencies a main effect was found for practice (with the practice groups taking less time to answer the post-test items than the no practice group) and for visual elaboration (with the students who were enrolled in the course with animated graphics took less time to answer the post-test items than the students who had static graphics or no graphics in their course). No main differences between static graphic and no graphics, and between practice questions and practice with a simulation could be detected. It is interesting to observe that the post-test scores on the whole were sensitive enough to differentiate between practice and no-practice and, moreover, between animated graphics, and static or no-graphics.

In Rieber (1990) an identical procedure was followed, only with younger students; a similar post-test was used, and therefore the results can be compared. In this study the correctness scores of the post-test did not show significant differences between the experimental groups. The latency data, however, revealed a main effect for practice: the groups that received the simulation or the practice questions had lower response times than the no-practice group.

Study (authors)	Assessment method
White, 1984	verbal items, some with open answers, some with 3 response alternatives requiring predictions of situations, phrased in one of the following two ways: (1) What would happen if...? (2) How could one achieve?
White, 1993	13 items with two or three response alternatives, most with visuals depicting the situation, requiring predictions of situations
Kieras, 1993	3-choice items, electronically administered, involig a picture of a electronic circuit, a perturbing event, and the answers were a choice of a voltage that either increased, decreased, or stayed the same, both correctness and latency data were collected.
Rieber et al., 1990	19 multiple choice items, electronically administered, half of the items of test were all verbal, half also contained an accompanying visualisation, and half of the items with visualisation contained an animation, the items had 5 response alternatives (1 answer, 4 distractors) and required predictions from the students, resulting in correctness and latency data; latency was measured as time (in seconds) taken by students to press the key of the right answer, once prompted to do so, the prompt was the final text for each item (p. 48)
Rieber, 1990	26 items, see for details Rieber et al., 1990

Figure 2: Details of Assessment Methods of Studies in Which Predictions With or Without Time Pressure Were Applied

No effects of graphics were detected in this study. Figure 2 gives the details of the assessments of the reviewed studies.

White's work shows that asking learners to make predictions is a good way to assess knowledge acquired with simulation environments. This concept is also applied in Rieber's items in the broader context of science education. In addition, both of Rieber's studies show that "the combination of performance and latency data provide a more complete understanding of learning than does either data source alone" (1991, p. 323). In this work we follow Rieber and we conclude that latencies may reveal interesting qualities of knowledge. More specifically, we believe that items requiring quick predictions or *quick perceptions of meaningful situations* may tell us something about discovery learning and the intuitive quality of the resulting knowledge.

WHAT-IF Format and Procedure

Based on our definition of intuitive knowledge as being characterised by a *quick perception of meaningful situations* and on the assessment techniques used in other studies we can present our proposed test format by following the key-words in our definition:

Quick: We decided to include response time of the items as an important indicator of the degree of intuitive quality of conceptual knowledge. This is because we believe that intuitive quality makes access to the knowledge more efficient.

Perception: In the item format we propose, we point to the importance of perception and contrast this with the emphasis of many other traditional tests on verbalisation. We therefore use pictures, graphical, or diagrammatic presentations accompanied by the minimal necessary textual information. We have two interrelated reasons for using rich graphical displays in the items. The first is that the intuitive quality is shaped by experiential processes which are fed by perception elicited by rich graphical displays. Reading many words, on the other hand, may elicit reflective thinking in which we are not primarily interested here. Secondly, by using rich graphical displays in the items we keep the verbal component of the items as low as possible. This low verbal level of the items is in line with the idea that the knowledge we want to measure, because of its intuitive quality, is hard to verbalise.

Meaningful: The general goal of learning with simulations is to discover the relations between variables. The learners are supposed to explore the material by changing variables and looking at the consequences. The questions consist of situations (see *Situations*) in which values of variables are given; a value is then changed and a new situation is to be predicted. The nature of the items is closely linked to the general goal of learning with simulations and we therefore assume that the situations described in the items should be meaningful to the learners.

Situations: The items consist of a question and possible answers. In the question part a description of a situation is given along with a change in that situation. In the answer parts descriptions of possible predicted situations are given. In other words, an item contains a situation, an action, and possible post-situations (or, in other words: a condition, an action, and predictions). The condition-part is described by variables which

are given a value, in the action-part a value of one variable is changed, and in the prediction-part possible new values of one of the variables of the condition-part are displayed. Furthermore, the pictures, graphs, or diagrammatic presentations used to describe the situations in the items are the same as those used in the simulation environment

We call the proposed item format, the WHAT-IF item format. In the WHAT-IF task, conceptual knowledge is presented in the form of conditions, actions and predictions. The conditions and predictions are the states or situations in which the system can be involved, and they can be displayed in a drawing of the system, or in a graph or tabular form, accompanied by a minimal amount of text. In addition, predictions concerning one variable can be presented by one word or number. The action, or the change of a variable within the system, is presented in text and/or graph. The items of the task are kept as simple as the domain permits. Items can be qualitative or quantitative, but in the case of quantitative items, no calculations are needed for reaching a correct answer. Knowledge about the specific nature of the relation between the variables is sufficient.

An important aspect of the procedure is that learners are not only asked to give the correct answer, but they are also required to do so as quickly as they possibly can. In this, the WHAT-IF format and procedure clearly differ from the format and procedures used in traditional multiple-choice tests. Like most multiple-choice tests, the WHAT-IF tests can be administered to classes of students in a relatively short period of time, and they are easily scored.

In the next sections we present two pilot studies in which a first evaluation of the WHAT-IF format was performed. Sample items are displayed within the sections of the pilot studies (Figures 3 and 4). The items had a three-answer format. The task was computer administered. The moment learners clicked the answer of their choice, the item disappeared from the screen and the next item popped-up. Latency was measured as the time (in seconds) learners needed to read and respond to the item.[4] In the pilots, this time corresponded to the time each item was displayed on the screen. Learners could not go back to previously answered items.

A First Validation of the WHAT-IF Format

General Procedure

The first validation studies we conducted followed a pre-test post-test design,[5] and parallel versions of WHAT-IF tests were applied as pre-and as post-tests to every learner. Apart from the WHAT-IF test, the learners were also administered a definitional knowledge test, asking for more traditional, verbalisable knowledge. The definitional knowledge test consisted of three-answer items and aimed to measure conceptual knowledge of facts and definitions of a declarative quality. The same definitional test was given both as pre- and as post-test. The correctness scores on the definitional tests were compared with the correctness scores and the completion times of the WHAT-IF tests.

Apart from gathering data on knowledge acquired, we registered all the actions learners performed while interacting with the simulation. This provided us with data on the use of the simulation and the supportive measures that were present. These process

measures were taken into account in the validation of the WHAT-IF format. The correctness scores and the completion times of the WHAT-IF tests were correlated with the number of runs, and the frequency of use of the several types of support.

The first learning environment dealt with the domain of elastic collisions and was called "Collision"; in the second environment harmonic oscillations were treated, and this learning environment was called "SETCOM". (A full description of the Collision study can be found in de Jong et al., 1995; the SETCOM study is fully described in van Joolingen, van der Hulst, Swaak, & de Jong, 1995).

Study 1: Collision

Domain of the Study and the WHAT-IF Items

The central topic of the Collision learning environment involves the physics topic of elastic collisions. More specifically, the rules behind colliding objects of different mass and velocity, and the concept of "centre of mass" are addressed. Furthermore, both kinetic and potential energy approaches are part of the simulation environment. For the evaluation of Collision two parallel WHAT-IF tests, each consisting of 37 items, were developed. The parallel versions of the tests were created in such a way that a one-to-one mapping existed between the WHAT-IF pre-test and WHAT-IF post-test items (i.e., item 1 of the pre-version corresponded with item 1 of the post-version, item 2 of the pre- with item 2 of the post-version, etc.). Parallel items covered the same content and had a similar difficulty level. During test development the items were reviewed to check whether the guidelines for the WHAT-IF format were interpreted correctly and to see how the guidelines were applied in this particular domain.

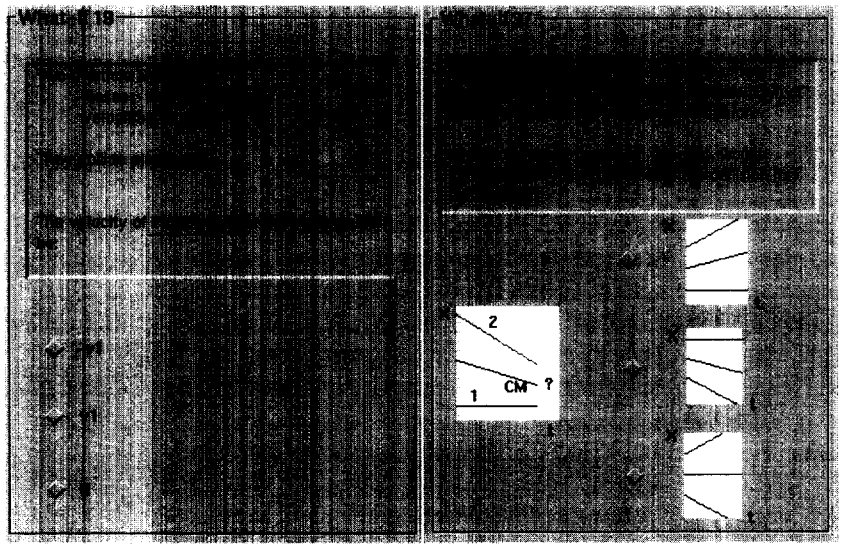

Figure 3: Two Examples of WHAT-IF Test Items Used in the Collision Study. (In the item on the left predictions are represented by one number/symbol. In the right hand side item the predictions are displayed by graphs. The items in the figure are translated from Spanish.)

After test development the tests were reviewed by an (independent) domain-pedagogical expert for content validity (i.e., completeness and accuracy) and readability (are the words used sufficiently easy, and is the number of words used sufficiently low for the type of items?). Two examples of WHAT-IF items are displayed in Figure 3.

Subjects

Students (N = 46) were selected from a population of the first year university level and came from the Mathematics/Computer science and Biology departments. Both types of students have comparable high school backgrounds, but the students from Mathematics/Computer-science passed through a selection process previous to be allowed to study Mathematics/Computer science.

Procedure

The Collision WHAT-IF pre- and post-tests were administered to 46 students before and after they worked, for about an hour, with the Collision simulation environment. Before the students started to work on the WHAT-IF tests they completed a definitional knowledge test, which consisted of 36 three-answer items. The same definitional test was administered both as pre- and as post-test version. The session ended after completion of the WHAT-IF post-test.

WHAT-IF Test Results

The average number of correctly answered items on the WHAT-IF pre-test was 19, with an *SD* of 4. On the WHAT-IF post-test the number correct scores had a mean of 23, with a SD of 5. The average time of completion of the WHAT-IF pre-test was 1211 seconds, with a *SD* of 283. For the WHAT-IF post-test the average completion time was 802 seconds, and the *SD* was 250.

No trade-off between correctness and speed was found. The correlations between answer time and correctness had a value of $r = -.17$, $p > .10$ when computed within students across the WHAT-IF pre-test items, a value of $r = -.23$, $p > .10$ when computed within students across the WHAT-IF post-test items, a value of $r = .02$, $p > .10$ when computed within WHAT-IF pre-test items across students, and finally a value of $r = -.27$, $p > .10$ when computed within WHAT-IF post-test items across students.

Repeated measurement analyses on the WHAT-IF test scores showed both a significant within-subject effect of number of correct items ($F_{1, 43} = 36.34$, $p < .05$) and a significant within-subject effect of the test completion times ($F_{1, 43} = 125.09$, $p < .05$).

Relationship Between the WHAT-IF Test Scores and Definitional Test Score

The definitional knowledge test was administered in the same form as the pre- and as post-tests. It consisted of 36 multiple choice items with three alternative answers each. A repeated measurement analysis showed a significant within-subject effect of number of correct items ($F_{1, 43} = 37.56$, $p < .05$). In Table 1 the correlations between the WHAT-IF post-test scores and the definitional test correctness scores are given.

Table 1: Correlations Between the WHAT-IF Post-Test Scores and the Definitional Knowledge Test Scores

Post-test scores	Definitional post-test	WHAT-IF post-test completion times
WHAT-IF correctness scores	.80 ($p < .01$)	.02 ($p > .10$)
WHAT-IF completion times	-.11 ($p < .10$)	

The pattern that emerges from this analysis reveals two clusters. The first consists of the definitional test and WHAT-IF correctness, the second is the WHAT-IF time aspect.

Relationship Between the WHAT-IF Test Scores and Several Process Measures

We registered all the actions students performed while interacting with the simulation. This provided us with data on the use of the simulation and the supportive measures (explanations, assignments, and model progression) that were present. Due to technical circumstances resulting in loss of logfiles, it was not possible to collect the interaction data of all the students. In the subsequent analyses all interaction data of 34 subjects were used. All data presented on interaction behaviour should be seen in this context.

As all the learners used the support measures available to them to a considerable extent, not much variation between subjects existed. However, some interesting relations were found between use of assignments and scores on the WHAT-IF post-test. Table 2 displays the WHAT-IF post-test scores, and number of explanations and assignments used.

Table 2: Correlations Between the WHAT-IF Post-Test Scores and Number of Assignments and Explanations Used

Post-test scores	Number of assignments	Number of explanations
WHAT-IF correctness scores	.51 ($p < .05$)	.15 ($p > .10$)
WHAT-IF completion times	-.35 ($p < .10$)	.11 ($p > .10$)

Table 2 shows a significant correlation between the number of assignments and the correctness scores of the WHAT-IF post-test, and non-significant (at the α level = .05) but moderate negative correlation between the numbers of assignments and the completion times of the WHAT-IF post-test. These figures indicate that a higher number of assignments is associated with a higher correctness score on the WHAT-IF post-test, and suggest that

the more assignments learners have used, the lower their completion times of the WHAT-IF post-test are.

Table 3: Correlations Between the WHAT-IF Post-Test Scores and Number of Runs Used

Post-test scores	Number of runs
WHAT-IF correctness scores	.13 ($p > .10$)
WHAT-IF completion times	-.02 ($p > .10$)

The figures in Table 3 show that none of the correlations reaches a level of significance below .05. We, therefore, can not conclude that a higher number of runs is associated with higher WHAT-IF correctness scores or lower WHAT-IF completion times.

Discussion

The overall picture that emerges is quite satisfactory. A knowledge gain is found and established as both an increase in correct WHAT-IF items and a decrease in WHAT-IF test completion times. Furthermore, the items seem neither too difficult nor too easy. Importantly, no trade-off is found between the correctness and the completion time of items. There even seems an indication of the reverse: the quicker an item is answered, the greater the chance it is correct. The suggested direction is clearly in line with our ideas on the intuitive quality of conceptual knowledge as the quick perception of meaningful situations.

However, a significant gain is not only found on the WHAT-IF tests but also on the definitional tests, and more essentially, a strong correlation is found between the correctness scores of the WHAT-IF post-test and the scores of the definitional post-test. The high correlation makes it more difficult to maintain that the tests measure different aspects and qualities of the domain knowledge. Yet, the correlation between the completion times of the WHAT-IF post-test and the scores on the definitional test is not found to be significant.

Finally, the moderate correlations between the number of assignments used and the WHAT-IF post-test scores may suggest that the more assignments learners have utilised, the higher their WHAT-IF post-test correctness scores, and the lower their completion times of the WHAT-IF post-test. These figures correspond nicely with our assumption that assignments would encourage a more experiential mode of learning and have a beneficial effect on students' intuitive knowledge.

Study 2: SETCOM

Domain of the Study and the WHAT-IF Items

The learning environment SETCOM deals with the principles of one-dimensional oscillatory motion. It incorporates three kinds of oscillatory motion: free oscillatory motion

without friction, damped motion, and forced oscillatory motion. Moreover the concepts of *sub- and supercritical damping*, *resonance* and the characteristic equation are addressed. For the evaluation of SETCOM two parallel WHAT-IF tests each consisting of 34 items were developed. The tests consisted of both qualitative (21 items) and quantitative items (13). Two examples of WHAT-IF items are displayed in Figure 4.

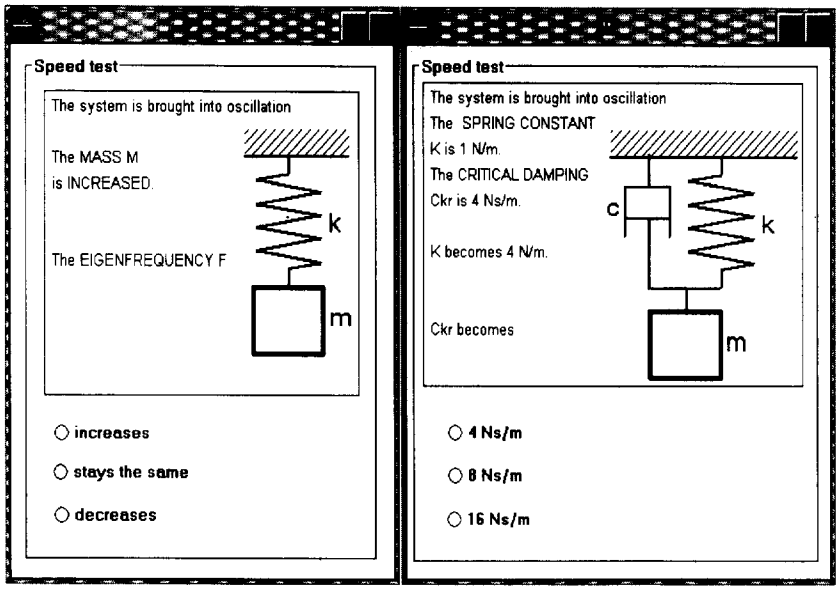

Figure 4: Two Examples of WHAT-IF Test Items Used in the SETCOM Study. (On the left a qualitative item, on the right a quantitative item. The items in the figure are translated from the Dutch.)

Subjects

Twenty-eight subjects participated in the study. They were undergraduate students in Mathematics/Computer science, Chemistry, and Psychology. All subjects had studied physics, including classical mechanics, at A level; only the Mathematics/Computer science students had a background in the theory of complex numbers. Most subjects were not familiar with the domain of oscillatory motion.

Procedure

The procedure in this pilot is the same as in the previous study, except that here the 28 learners worked for about two hours with the simulation environment.

WHAT-IF Test Results

The average number correctly answered items on the WHAT-IF pre-test was 11, with an *SD* of 4. On the WHAT-IF post-test the number correct scores had a mean of 16,

with a *SD* of 3. The average time to complete the WHAT-IF pre-test was 781 seconds, with a *SD* of 215. For the WHAT-IF post-test the average completion time was 655 seconds, and the *SD* was 227.

We did not find a trade-off between correctness and speed. The correlations found between answer time and correctness had a value of $r = .00$, when computed within students across the WHAT-IF pre-test items, a value of $r = .00$, when computed within students across the WHAT-IF post-test items, a value of $r = -.16$, $p > .10$ when computed within WHAT-IF pre-test items across students, and finally a value of $r = -.19$, $p > .10$ when computed within WHAT-IF post-test items across students.

Repeated measurement analyses were performed on the WHAT-IF test scores. The analyses showed both a significant within-subject effect of number of correct items ($F_{1, 26} = 18.34$, $p < .01$) and a significant within-subject effect of the test completion times ($F_{1, 26} = 12.93$, $p < .01$).

Relationship Between the WHAT-IF Test Scores and Definitional Test Scores

The definitional knowledge test was administered in the same form, both as pre- and as post-test. It consisted of 30 multiple choice items with three alternative answers each. A repeated measurement analysis on the definitional test-scores showed no significant within-subject effect of number of correct items ($F_{1, 25} = 1.77$, $p > .10$). Table 4 displays the correlations between the post-test scores of the WHAT-IF tests and the definitional knowledge test scores.

Table 4: Correlations Between the Post-Test Scores of the WHAT-IF Tests and the Correctness Scores of the Definitional Knowledge Test

Post-test scores	Definitional post-test scores	WHAT-IF post-test completion times
WHAT-IF correctness scores	.42 ($p < .05$)	-.01 ($p > .10$)
WHAT-IF completion times	.19 ($p > .10$)	

Here we see the same pattern as in the previous study.

Relationship Between the WHAT-IF Test Scores and Several Process Measures

All the learners used the support measures available to them to a considerable extent and therefore variation between subjects was small. Still, some correlations worth mentioning were found. Table 5 displays the correlations.

The fact that the correlation reaches α level = .05 indicates that the higher number of explanations used, the higher the WHAT-IF post-test completion times. The negative correlation of -.38 suggests that the more assignments used, the lower the WHAT-IF post-test completion times. This same negative correlation was found in the previous study.

Table 5: Correlations Between the WHAT-IF Post-Test Scores and Number of Assignments and Explanations Used

Post-test scores	Number of assigments	Number of explanations
WHAT-IF correctness scores	.00 ($p > .10$)	-.19 ($p > .10$)
WHAT-IF completion times	-.38 ($p < .10$)	.38 ($p < .05$)

The figures in Table 6 show that none of the correlations reaches a level of significance below .05. Like in the Collision study, we can, therefore, not conclude that a higher number of runs is associated with higher scores.

Table 6: Correlations Between the WHAT-IF Post-Test Scores and Number of Runs Used

Post-test scores	Number of runs
WHAT-IF correctness scores	.11 ($p > .10$)
WHAT-IF completion times	-.27 ($p > .10$)

Discussion

This study generally reveals that the performance on the post-test measures is quite low. More specifically, the correctness scores on the WHAT-IF pre-test are at chance level, and the correctness scores on the WHAT-IF post-test are still near chance level.

Nevertheless, both a significant increase in correctness scores and a significant decrease in completion times of the WHAT-IF tests are established, and furthermore, no trade-off between correctness and completion times is found. Like in the Collision study, there even seems to be an indication of the reverse, which fits well with our hypotheses.

In contrast to the gain in WHAT-IF test scores, no increases in definitional knowledge test scores are attained. Another interesting finding is the relation between the number of assignments used and the results on the WHAT-IF post-test. A moderate negative - non-significant at the α level =.05 - correlation is found between the number of assignments and the completion times of the WHAT-IF post-test. This may suggest that the more assignments the learners utilised, the faster they completed the WHAT-IF post-tests. The reverse is established for the number of explanations: the more explanations are looked up, the longer the completion times of the WHAT-IF post-test. The rationale put forward here is that explanations foster a reflective mode of learning, resulting in higher response times of WHAT-IF items, whereas assignments invite learners to follow a experiential learning mode which is assumed to improve the intuitive quality of knowledge, as reflected in lower response times of WHAT-IF items.

Conclusion

An overall conclusion from our study is that the two tests constructed with WHAT-IF items were able to tap an improvement in learning in the two pilot studies. The WHAT-IF tests seemed furthermore capable of assessing the positive impact of the use of assignments (both studies), and the less favourable influence of explanations (the SETCOM study). Moreover, there is some evidence in the data that the more quickly an item is answered, the greater the chance that it is correctly answered. All these figures correspond to our conceptions on the intuitive quality of knowledge, and to our ideas about which circumstances promote the acquisition of knowledge with an intuitive quality. Therefore, we can say that the validity data are - as a first start - supportive. However, further studies are needed to validate the WHAT-IF item format and make the validity data more convincing. Some methods to improve the WHAT-IF tests are suggested. The first and foremost thing we plan to do is to perform item analyses including the contents of each individual WHAT-IF item. So far we have looked at the correctness and response times of the individual items, but we haven't related this information to the contents of the particular items. By doing so, we expect to arrive at more certain conclusions about which features of the items allow, for example, for faster correct responses, or which item characteristics prevent quick, meaningful responses or which features foster quick, consistent incorrect responses. We also hope to gain more insight into what a quick, intuitive answer constitutes, and what definitely not. Secondly, we propose a slight change in the procedure followed in the administration of the tests. In future applications of the WHAT-IF tests, learners will receive some practice items before starting the real test. During practice learners are allowed to ask questions concerning the procedure, and at the same time get used to the interface, mouse, etc. Another step we are considering, is to find out whether an Item Response model (IRT model), combining correctness and latency, fits our data.

We would like to conclude with a more detached view on the whole venture. The enterprise of evaluating the latest conceptions on simulation based learning environments, with the idea of tapping a certain quality of knowledge by means of a new testing format, which is then to be validated both by assessment techniques applied for the first time in this context, as well as by new methods of recording learning processes, is doomed to be risky. The rationale for giving it a go anyway is that technology is rapidly realising new ways of instruction, while, in our view, assessment is lagging behind. Assessment can profit from technology as well, and moreover, this rapidly evolving technology, in its turn, stands to gain substantially from the output of proper measurement. Therefore, as a general conclusion we think that our data show that designing new types of knowledge tests in relation to new types of (technology supported) learning is a promising way to follow.

Notes

1. We would like to thank our colleagues Wouter van Joolingen, Jules Pieters, and Cees Glas (University of Twente) for providing us with fruitful ideas for this study. The pilot studies we

report were conducted in co-operation with Ernesto Marton, Jose-Miguel Zamarro, Francisco Esquembre (University of Murcia), Anja van der Hulst (University of Amsterdam) and Wouter van Joolingen (University of Twente). Part of the work presented here was conducted in the SMISLE (D2007) and SERVIVE (ET 1020) projects, both sponsored by the EC in its Telematics programmes under DG XIII.

2. The SMISLE system is an authoring environment for developing Multimedia Integrated Simulation Learning Environments.
3. Complex domains refer in this context to knowledge rich domains and can be contrasted with basic psychological research in which among others, simple recognition and recall tasks are studied. In this kind of research the use of latencies is widespread in connection with implicit knowledge (e.g., Reber, 1989, 1993; Lewicki & Hill, 1989).
4. This is unlike the procedure followed by Kieras and Rieber who distinguished between reading time of the item and response time. Here, it was decided not to split reading and response time, as some learners may anticipate answers while reading the items, and others may not, inducing unintended effects.
5. The students in the pilot studies were divided across experimental conditions (in Study 1 we created three, and in Study 2, two groups) which differed with respect to the instructional measures added to the simulation. In this article only analyses that were performed across the conditions are reported.

References

Anderson, J.R. (1987). Skill acquisition: Compilation of weak-method problem solutions. *Psychological Review, 94 (2),* 192-210.

Berkum, J.J., van A., & Jong, T., de (1991). Instructional environments for simulations. *Education & Computing, 6,* 305-359.

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

Berry, D.C., & Broadbent, D.E. (1988). Interactive tasks and the implicit-explicit distinction. *British Journal of Psychology, 79,* 251-272.

Berry, D.C., & Broadbent, D.E. (1990). The role of instruction and verbalization in improving performance on complex search tasks. *Behaviour and Information Technology, 9* (3), 175-190.

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.

Brown, D.E. (1993). Refocusing core intuitions: A concretizing role for analogy in conceptual change. *Journal of Research in Science Teaching, 30* (10), 1273-1290.

Chase, W.G., & Simon, H.A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55-81.

Chmiel, N., & Tattersall, A. (1991). Implicit and explicit knowledge: Display factors in the control of complex processes. *SAPU Memo No. 1240.* MRC/ESRC Social and Applied Unit: University of Sheffield.

Fischbein, E. (1987). *Intuition in science and mathematics.* Dordrecht, The Netherlands: Reidel.

Glaser, R., Raghavan, K., & Schauble, L. (1988). Voltaville, a discovery environment to explore the laws of DC circuits. *Proceedings of the ITS-88* (pp.61-66). Montreal, Canada.

Groot, A.D., de (1965). *Thought and choice in chess.* The Hague: Mouton.

Hayes, N.A., & Broadbent, D.E. (1988). Two modes of learning for interactive tasks. Cognition, 28, 249-276.

Jong, T., de, & Ferguson-Hessler, M.G.M. (1996). Types and qualities of knowledge. *Educational Psychologist, 31,* 105-113.

Jong, T., de, & Joolingen, W.R., van (1995). The SMISLE environment: Learning and design of integrated simulation learning environments. In P. Held & W.F. Kugemann (Eds.), *Telematics for education and training* (pp. 173-187). Amsterdam: IOS Press.

Jong, T., de, & Joolingen, W.R., van (in preparation). *Discovery learning with computer simulations.*

Jong, T., de, Joolingen, W.R., van, Pieters, J.M., Hulst, A., van der, Hoog, R., de (1992). *Instructional support for simulations: Overview, criteria and selection.* DELTA project SMISLE, Deliverable D02. University of Twente, Department of Education.

Jong, T., de, Martin, E., Zamarro J-M., Esquembre, F., Swaak, J., & Joolingen, W.R., van (1995, April). *Support for simulation-based learning; the effects of assignments and model progression in learning about collisions.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Joolingen, W.R., van, Hulst, A., van der, Swaak, J, & Jong, T., de (1995). *Support for simulation-based learning; the effects of model progression in learning about oscillations.* SMISLE project, deliverable D24a. Enschede: University of Twente.

Kieras, D.E. (1993). Learning schemas from explanations in practical electronics. In S. Chipman & A.L. Meyrowitz (Eds.), *Foundations of knowledge acquisition: Cognitive models of complex learning* (pp. 83-117). Boston/Dordrecht/London: Kluwer.

Larkin, J.H.,& Simon, H.A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11,* 65-99.

Leutner, D. (1993). Guided discovery learning with computer-based simulation games: Effects of adaptive and non-adaptive instructional support. *Learning and Instruction, 3,* 113-132.

Lewicki, P., & Hill, T. (1989). On the status of nonconscious processes in human cognition: Comment on Reber. *Journal of Experimental Psychology: General, 118,* 239-241.

Lindström, B., Marton, F., Ottosson, T., & Laurillard, D. (1993). Computer simulations as a tool for developing intuitive and conceptual understanding in mechanics. *Computers in Human Behavior, 9,* 263-281.

Myers, C., & Davids, K. (1993). Tacit skill and performance at work. *Applied Psychology: An International Review, 42* (2), 117-137.

Norman, D.A. (1993). *Things that make us smart: Defending human attributes in the age of the machine.* New York: Voyager.

Reber, A.S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General, 118,* 219-235.

Reber, A.S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious.* New York: Oxford University Press.

Rieber, L.P. (1990). Using computer animated graphics in science instruction with children. *Journal of Educational Psychology, 82,* 135-140.

Rieber, L.P. (1991). Animation, incidental learning, and continuing motivation. *Journal of Educational Psychology, 83* (3), 318-328.

Rieber, L.P., Boyce, M.J., & Assad, C. (1990). The effects of computer animations on adult learning and retrieval tasks. *Journal of Computer Based Instruction, 17* (2), 46-52.

Rieber, L.P., Smith, M., Al-Ghafry, S, Strickland, B., Chu, G., & Spahi, F. (1995). *The role of meaning in interpreting graphical and textual feedback during a computer-based simulation.* Paper presented at the Annual Meeting of the American Educational Research Association, San Fransisco.

Shute, V.J., & Glaser, R. (1990). A large-scale evaluation of an intelligent discovery world: Smithtown. *Interactive Learning Environments, 1,* 51-77.

Thomas, R., & Hooper, E. (1991). Simulations: An opportunity we are missing. Journal of Research on Computing in Education, 23(4), 497-513.

Wagner, R.K., & Sternberg, R.J. (1986). Tacit knowledge and intelligence in the everyday world. In R.J. Sternberg, & R.K. Wagner (Eds.), *The nature and origin of competence in the everyday world.* Cambridge: Cambridge University Press.

Wagner, R.K., & Sternberg, R.J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 49* (2), 436-458.

Westcott, M.R. (1968). *Towards a contemporary psychology of intuition.* New York: Holt, Rinehart & Winston.

White, B.Y. (1984). Designing computer games to help physics students understand Newton's laws of motion. *Cognition and Instruction, 1* (1), 69-108.

White, B.Y. (1993). Thinker tools: Causal models, conceptual change, and science education. *Cognition and Instruction, 10* (1), 1-100.

White, B.Y., & Frederiksen, J.R. (1990). Causal model progressions as a foundation for intelligent learning environments. *Artificial Intelligence, 42,* 99-157.

The Authors

JANINE SWAAK is currently working on her PhD thesis on assessment of the outcomes and processes of simulation-based learning, at the University of Twente, the Netherlands. She has an MA degree in research methods and statistics and cognitive psychology, Faculty of Psychology, University of Amsterdam.

TON DE JONG studied cognitive psychology at the University of Amsterdam and received a PhD from Eindhoven University of Technology on the topic of problem solving and knowledge representation in physics for novice students. He joined Delft University of Technology; later he became researcher, then senior researcher, at both the University of Amsterdam and Eindhoven University of Technology. He has been Associate Professor at the University of Twente since 1992.