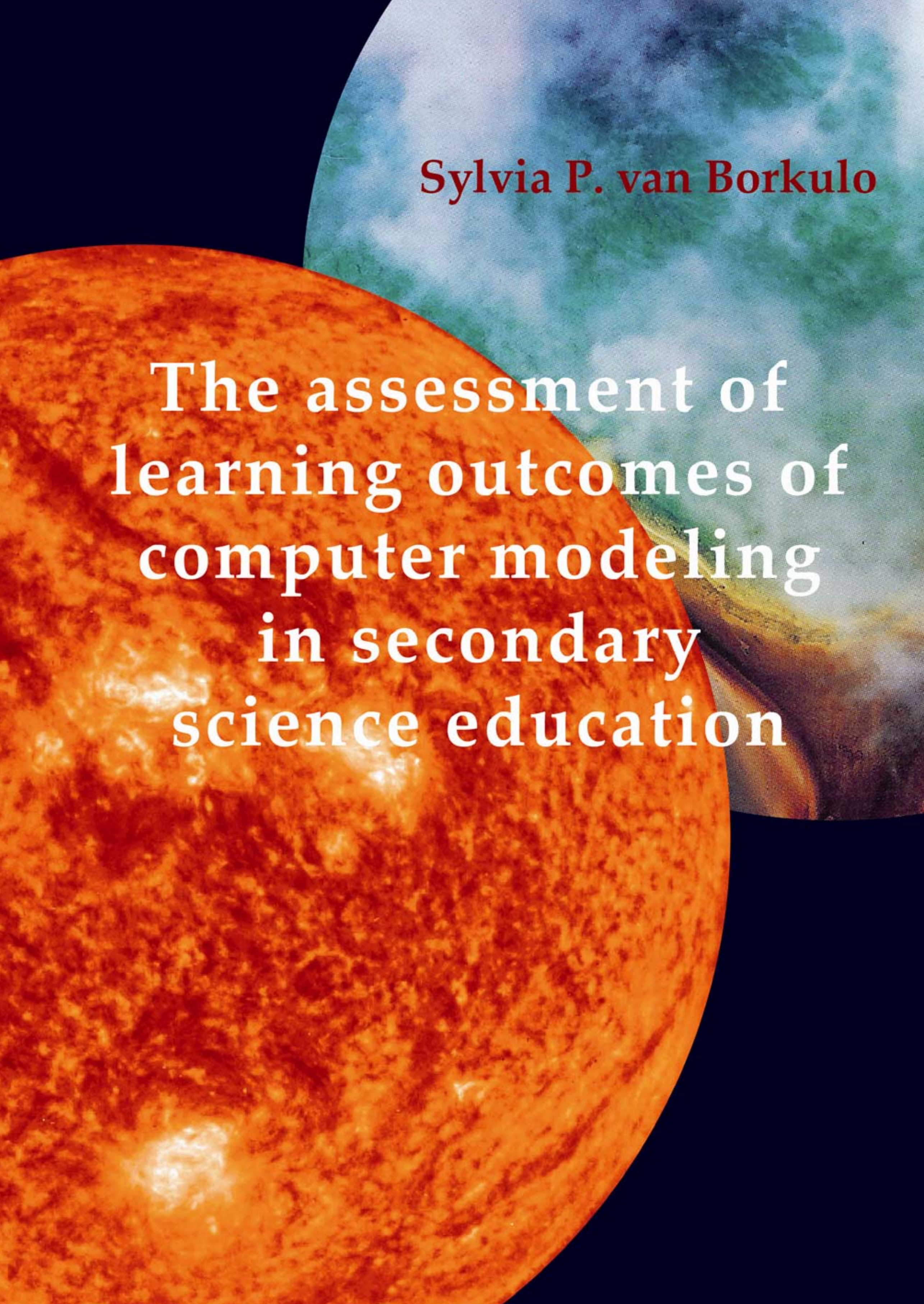


Sylvia P. van Borkulo

The assessment of learning outcomes of computer modeling in secondary science education

The assessment of learning outcomes of computer modeling in secondary science education

Sylvia P. van Borkulo

Doctoral committee

Chair:	Prof. dr. H.W.A.M. Coonen
Promotoren:	Prof. dr. W.R. van Joolingen
	Prof. dr. A.J.M. de Jong
Assistent-promotor:	Dr. E.R. Savelsbergh
Members:	Prof. dr. C.A.W. Glas
	Prof. dr. K.P.E. Gravemeijer
	Prof. dr. E. Klieme
	Prof. dr. J.M. Pieters
	Dr. A. Weinberger

The research reported in this thesis was funded by the Netherlands Organization for Scientific Research (grant number 411-03-133).

This research was carried out in the context of the Interuniversity Centre for Educational Research.

ISBN: 978-90-9024394-8

Cover design: Jan-Maarten Luursema
Print: PrintPartners Ipskamp, Enschede, The Netherlands

THE ASSESSMENT OF LEARNING OUTCOMES OF COMPUTER MODELING IN SECONDARY SCIENCE EDUCATION

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. H. Brinksma,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 26 juni om 13.15 uur

door

Sylvia Patricia van Borkulo
geboren op 27 april 1972
te Amsterdam

Dit proefschrift is goedgekeurd door de promotoren:
Prof. dr. W.R. van Joolingen
Prof. dr. A.J.M. de Jong

en assistent-promotor:
dr. E.R. Savelsbergh

Dankwoord

Het proefschrift dat in uw handen ligt, is het resultaat van samenwerking met vele mensen. In dit dankwoord wil ik graag enkele mensen in het bijzonder bedanken.

Het onderzoek werd begeleid door een inspirerend driemanschap. Ton, ik wil je danken voor de gelegenheid die je me hebt gegeven om in jouw onderzoeksgroep onderzoek te doen naar een onderwerp dat mijn hart heeft gestolen. Elwin, ik dank jou voor het stellen van al die vragen die me steeds een stap verder hielpen. Wouter, bedankt voor het vertrouwen dat je altijd in me stelde en de steun die je me gaf in raad en daad.

De studies beschreven in dit proefschrift hadden niet tot stand kunnen komen zonder de bereidwillige inzet van scholen, docenten en leerlingen. Ik dank de docenten Benno Berendsen van het Bonhoeffer College, Tjalling Visser, Jaap Andela en René Mondeel van CSG Het Noordik, de technische ondersteuning op beide scholen en alle leerlingen die hebben deelgenomen aan mijn experimenten. Jullie inzet en enthousiasme hebben mijn onderzoek tot een succes gemaakt.

Mijn collega's van de afdeling Instructietechnologie dank ik voor de prettige samenwerking. Jakob, Elske, Hannie, Yvonne, Wout, Bas, Marleen, Sarah en Mieke, bedankt voor jullie onmisbare hulp bij de experimenten. Anjo and Wout, bedankt voor de gezellige avondjes promotieliedjes rijmen. Ik heb vreselijk gelachen. Jammer dat ik de voorbereidingen voor mijn eigen liedje moet missen... Mijn kamergenoten Cornelise, Irina en Tim, bedankt voor de broodnodige afleiding en gezelligheid. Petra, bedankt voor je advies om *niet* naar de bibliotheek te gaan.

Buiten mijn werk heb ik steun gehad van vele mensen. Mijn huisgenoten in 't Piepke: bedankt voor jullie 'begeleiding'! Alle vrienden in het verre Westen die ik veel te lang heb moeten missen, ik hoop jullie gauw weer te zien. Sandra en Puccini, bedankt voor alle energie die jullie me gegeven hebben.

Tot slot, Harm, duizend dank!

Sylvia van Borkulo,
Juni 2009

Table of contents

Chapter 1 Introduction

Education is changing, in the Netherlands as well as elsewhere around the world. Many of these changes have been initiated because of a perceived need for new learning outcomes. An example is the reform of science education that emphasizes a general shift toward realistic tasks and higher-order learning goals, such as the acquisition of scientific literacy, inquiry skills, and a hands-on and minds-on mentality (van Driel, Beijaard, & Verloop, 2001). For each element of this reform, critics have been eager to deny that the intended outcomes have been, or even could have been achieved. Unfortunately, the debate between reformers and their critics all too often remains undecided, due to a lack of rigorous evidence to convince the other side. Apparently, higher-order learning outcomes can be very hard to assess.

Inquiry learning provides an example of a newly introduced educational means for achieving the new learning outcomes. In inquiry learning, students are expected to increase their scientific literacy by engaging in activities such as experimenting and constructing knowledge 'like scientists do' (Van Joolingen, De Jong, & Dimitrakopoulou, 2007). On the one hand, there is research suggesting that inquiry learning provides a motivating and engaging method of learning (Hanauer et al., 2006; Kuhn, Black, Keselman, & Kaplan, 2000; Lederman, Lederman, Wickman, & Lager-Nyqvist, 2007; Linn, Lee, Tinker, Husic, & Chiu, 2006; Schwartz, Lederman, & Crawford, 2004). On the other hand, there are also researchers and educators who maintain that direct instruction would be a much more effective approach (e.g., Kirschner, Sweller, & Clark, 2006). Klahr and Nigam (2004) compared ‘learning from direct instruction’ with ‘discovery learning’, using a learning task about experimental design, and found higher learning gains for the direct instruction learners. However, it must be noted that the implementation of direct instruction in Klahr and Nigam's research could equally well be regarded as scaffolded or guided inquiry learning. Interpreted in this way, the finding by Klahr and Nigam is a confirmation of the well-known finding that guided discovery is more effective than pure discovery (Mayer, 2004).

Although such terminological confusion must certainly be resolved before any meaningful comparisons can be made, the more pressing problem is the lack of suitable assessment tools for higher-order learning outcomes. Insofar as inquiry learning aims at new learning outcomes, such as generating hypotheses and evaluating experimental data, these outcomes won’t be detected by a traditional test aimed at reproducing and applying facts and formulae. It is quite plausible that the direct instruction approach will lead to better learning outcomes on such traditional tests, whereas the inquiry approach might do better on a test specifically aiming at the kinds of knowledge that are targeted by inquiry learning, e.g. intuitive knowledge

(Swaak & de Jong, 1996). In the context of today's quest for an evidence-based (Davies, 1999) choice between different forms of education, this leads to the conclusion that the development of appropriate assessment tools is vital.

A central aspect of inquiry learning is that the learners must develop their own models. In the Netherlands, one aspect of the reform in secondary science education is a larger curricular focus on the development of students' knowledge and abilities in the field of models and inquiry modeling (van Driel & Verloop, 2002). Here, we find similar problems with respect to how to instruct and how to assess the learning outcomes. Therefore, the focus of this dissertation is the assessment of learning outcomes of computer modeling in upper secondary education in the educational context of a modeling task. We limit ourselves in this to modeling *dynamic systems*, i.e., systems that autonomously change over time, as computer modeling is especially suitable for such domains.

1.1 Computer modeling of dynamic systems in education

When learning about a science topic, say, thermodynamics, a 'traditional' approach would involve the presentation of central concepts (heat, temperature, energy, entropy) and relations between concepts in the form of formulae (e.g., temperature = energy / heat capacity). Students are provided with exercises, which involve solving standard domain problems by applying these formulae and computing outcomes. Computer modeling offers an alternative form of instruction that asks learners to construct *executable models* of thermodynamic systems based on the central concepts and principles of the domain. Learners specify their models using a dedicated language and the computer can simulate the behavior of the modeled system as defined by the learners' models. So, instead of solving standard domain problems in a more or less algorithmic way, students' attention is shifted toward analyzing the domain itself in terms of its constituent concepts and using these concepts to construct a model.

In this dissertation, the construction and/or modification of executable computer models of dynamic phenomena (following Löhner, 2005) is taken as a defining characteristic of computer modeling. Learners in this context perform a scientific inquiry task that involves scientific reasoning processes, such as causal reasoning, considering implications of conditions or options (Jonassen, Carr, & Yueh, 1998), interpreting data, evaluating models, predicting, and explaining (Schwarz & White, 2005). Learners are assisted in these processes by the capability of the computer-based modeling tool to simulate the behavior of their models. Defining modeling in this way distinguishes it from other uses of the term that include working with a model that is given, for example in the form of a simulation or a formula (cf. exploratory modeling in Bliss et al., 1992). In so doing, we adopt the definitions used in earlier research on computer modeling in secondary education (e.g., Alessi, 2005; Bliss et al., 1992; Löhner, van Joolingen, Savelsbergh, & van Hout-Wolters, 2005; Manlove, Lazonder, & De Jong, 2006; Sins, 2006; Stratford, 1997).

Modeling is a form of active, generative learning in which knowledge construction is supported by creating and adapting a model (Forrester, 1994). This process is expected to lead to learning outcomes such as the ability to create and revise a model. Moreover, one may expect outcomes with regard to the learners' mental representation of the domain. In modeling complex systems the learner creates a model by translating internal ideas into an external representation. By confrontation with this external representation, or as an effect of evaluating the model, the learners' internal ideas, their mental model of the domain, may be adapted, leading to a cycle of revisions and tests of both the internal and external models.

Such learning outcomes from modeling are qualitatively different from the knowledge and skills developed in traditional instruction (Hestenes, 1996). Computer modeling may also reveal distinctive learning outcomes in comparison to the more closely related simulation-based learning, in which the simulation of a given model is used (De Jong & Van Joolingen, 1998), in the sense that modeling fosters the ability to create a model (Papert & Harel, 1991). Like modeling, simulation-based learning requires the design, performance, and interpretation of experiments, but no artifact is created.

Scientific reasoning, especially with dynamic systems, is difficult for a high school student (Jacobson & Wilensky, 2006) and even for university students (Hmelo-Silver, Nagarajan, & Day, 2002). For example, in the modeling of ocean waves students may have difficulty understanding that the energy of the wave moves and not the constituent parts of the water molecules (Wilensky & Resnick, 1999). In a complex dynamic system the behavior of the system as a whole may be different than the sum of the behavior of its parts. Other difficulties occur in the learning of scientific reasoning, such as problems in learning causal reasoning (Cronin & Gonzalez, 2007), systems thinking (Booth Sweeney & Sterman, 2000), and the effects of nonlinear relations over time (Milrad, Spector, & Davidson, 2003).

Scientific reasoning in modeling tasks is also complicated for students by the fact that modeling does not provide a standard procedure with determined outcomes. In modeling there can be multiple right answers and many ways to reach a solution, making it difficult to learn what to do and how to create a model. The possibility of multiple solutions introduces uncertainty and difficulty, because, in general, students want to know the ‘right answer’ or some kind of rule for getting the answer (Skemp, 2006). Despite the apparent difficulties of computer modeling, research has shown that students are able to overcome the problems in a modeling task (Stratford, Krajcik, & Soloway, 1998). It is claimed that through modeling, students acquire model-based scientific reasoning skills (Milrad et al., 2003), learn about the domain (Schecker, 1998), and gain insight into the behavior of (complex) dynamic systems in general (Forbus, 1996; Hogan & Thomas, 2001; Stratford et al., 1998; Wilensky & Resnick, 1999).

The above findings and claims are based mainly on observations and case studies, and different studies investigating these claims are not objectively comparable. To make a useful contribution to the development of effective modeling-based education, it is therefore necessary to clearly define the learning goals of modeling and to develop instruments to assess whether such learning goals are achieved. With such clearly defined learning goals and appropriate assessment instruments, one may demonstrate evidence for the specific effects of different modes of instruction (Davies, 1999). Unfortunately, many studies of the effects of modeling do not define the expected learning outcomes of modeling and lack appropriate measurement tools for modeling learning outcomes (Spector, 2000). This situation calls for a systematic investigation into learning outcomes of modeling. In this dissertation we aim to contribute to knowledge of computer modeling learning outcomes by performing such an investigation and by developing appropriate assessment methods. This leads to the following questions to be addressed: 1. what specific learning outcomes can be expected from modeling; 2. how can the specific learning outcomes be measured; and 3. what specific differences and similarities can be expected between the learning outcomes of modeling and of other modes of instruction?

1.2 Learning outcomes of computer modeling

The specific learning outcomes of a modeling task must be investigated in order to develop an appropriate assessment instrument for modeling-based learning. The analysis will focus on the types of outcomes that have been addressed by previous research as potential learning outcomes of modeling: scientific reasoning skills, conceptual knowledge of the domain involved, and insight into dynamic systems. Beyond these three areas, learning outcomes of modeling are also possible in the area of meta-knowledge such as ‘nature of science’, epistemological knowledge, and self-regulation (Guttersrud, 2006; Hogan &

Thomas, 2001; Manlove et al., 2006). Such learning outcomes are related to broader themes that go beyond computer modeling and the specific learning outcomes we are looking for in this research. Although these aspects of meta-knowledge are important enough to deserve attention in educational research, they are much more comprehensive than will emerge in a limited modeling task. Therefore, we will set aside these broad themes and focus on analyzing the cognitive learning outcomes of modeling. This analysis will enable us to derive a framework for the different types of learning outcomes and to operationalize the different types in test items.

1.2.1 Scientific reasoning skills

Computer modeling requires the performance of several processes of scientific reasoning, such as the generation of hypotheses, designing experiments to test them, and interpreting data. Therefore, performing a computer-based modeling task will form a suitable practice arena for acquiring and improving scientific reasoning skills. Scientific reasoning processes are considered to be higher-level thinking performances that are normally out of reach but that can be learned even at a lower secondary level with appropriate scaffolding (Fretz et al., 2002).

In defining the nature of these skills, Wells, Hestenes, and Swackhamer (1995) distinguish the processes of creating, evaluating, and applying models in concrete situations. A similar classification has been proposed by Löhner et al. (2005), who describe the modeling cycle as an iterative process that can be applied when developing and refining a model. The cycle is made up of the steps of orienting, hypothesizing, experimenting, modeling, and evaluating. In this terminology, the process of 'modeling' corresponds to creating variables and relations. Hypothesizing involves reasoning about the relations in the model, which requires applying the rules of a model. Experimenting is a process steered by the goal of evaluating the model and can be covered by the reasoning process of evaluating.

Stratford, Krajcik, and Soloway (1998) studied the thinking strategies that are best fostered by dynamic modeling and distinguish the processes of analyzing, relational reasoning, synthesizing, testing and debugging, and making explanations. More specifically, in Stratford et al.'s typology of thinking strategies, the scientific reasoning processes in which students engage during dynamic modeling are specified by the following aspects: identifying and creating factors or objects, making judgments, interpreting a model's behavior, drawing conclusions, creating and discussing relationships, predicting what should happen, viewing and evaluating the model as a whole, explaining relationships, stating evidence, and justifying an argument.

Overall, the processes for applying knowledge, creating variables and relations, and evaluating hypotheses and models appear to be central factors in scientific reasoning. These processes not only play a role in the modeling of dynamic systems, but also reflect a broader relevance as described by Bloom's Taxonomy (Bloom, 1956). This framework was developed to classify educational objectives in general; in the revision by Anderson and Krathwohl (2001) the cognitive processes of applying, creating, and evaluating are distinguished as higher-order learning gains.

International studies have reported on the assessment of scientific reasoning skills. For example, scientific reasoning is part of the scientific literacy assessment in the Programme for International Student Assessment (PISA) study (Harlen, 2001). In the PISA assessment, scientific literacy has been broadly defined by 'being 'at ease' with scientific ways of understanding things'. Scientific reasoning processes can be identified in the assessment of this generally defined ability, such as explaining relations, making predictions, identifying factors that influence a given outcome, and drawing and evaluating conclusions.

Scientific reasoning is also part of the assessment in the Trends in Mathematics and Science Study (TIMSS) (Mullis, Martin, Ruddock, Arora, & Erberber, 2005). In this study knowledge of science is assessed in the cognitive domains of knowing, applying, and reasoning. These domains cover processes such as recalling, relating, interpreting, explaining, predicting, evaluating, and drawing conclusions.

These examples of the assessment of scientific reasoning include scientific reasoning processes; however, these processes are assessed implicitly and integrated into a broader context, and lack a model-based context.

1.2.2 Conceptual domain knowledge

Apart from the acquisition of scientific reasoning skills, modeling promotes understanding of the science content (Stratford et al., 1998). Modeling has been studied as a means of learning in many domains, including water flow (Booth Sweeney & Sterman, 2000; Kainz & Ossimitz, 2002), thermodynamics (Forbus, Carney, Sherin, & Ureel, 2005; Löhner et al., 2005; Schecker & Einhaus, 2007), and ecosystems (Papaevripidou, Constantinou, & Zacharia, 2007; Stratford et al., 1998); biological topics such as plant growth (Ergazaki, Komis, & Zogza, 2005) and health and diet (Bliss et al., 1992); and logistics topics such as traffic (Bliss et al., 1992). These examples, covering a wide range of scientific disciplines, have in common that they represent dynamic phenomena and that the conceptual structure is crucial in understanding the domain. As computer modeling explicitly represents such structures, learners engaging in computer modeling are expected to be better able to acquire knowledge about them.

Domain knowledge can be both an input to and an output of the modeling process. In building a model the concepts in the domain get connected, both in the external model representation, and, as may be assumed, in the learner's mind. Thus, building an external model is assumed to support integrating information into mental models (Nersessian, 1999). The created computer model scaffolds and externalizes internal, mental models (Jonassen, Ströbel, & Gottdenker, 2005). The process of expressing internal mental models in external models leads to better understanding by requiring a precise definition of ideas and by providing opportunities to test the mental model (Doerr, 1996).

Conceptual domain knowledge involves not only isolated facts and relations, but also how these basic entities are connected. This means that the focus should be divided when assessing domain knowledge between the basic concepts and the way they compose the domain as a whole. Traditionally, conceptual knowledge is measured by knowledge tests asking for definitions of concepts and relations between them (Archbald & Newmann, 1988; Royer, Cisero, & Carlo, 1993). One potentially interesting enhancement of such assessment could be testing whether larger knowledge structures can be used in domain specific reasoning processes.

1.2.3 Insight into dynamic systems

By the nature of computer modeling, the dynamics of systems have a central focus in modeling tasks. Dynamic systems change over time autonomously, due to the fact that one or more components of the system are not in equilibrium. For instance, the weather above the Netherlands changes continuously due to differences in temperature and pressure, causing phenomena such as clouds, wind, and rain. A formal description of system components, their states and their relations, in the form of a computer model can help to gain insight into the structure and behavior of the system over time. In particular, simulating the model can demonstrate the system's dynamics and thereby support insight into dynamic behavior.

Dynamic behavior must be understood at two levels. On the level of the variables and relations, the dynamic aspects must be specifically defined in terms of the way variables change over time under the influence of other variables. The identification and naming of variables plays an important role in defining the model in a formal sense, for it determines the possibility of using the variables in formal relational causal reasoning. Reasoning with a relation between two variables has a form such as 'the greater the value of variable X, the greater will be the value of variable Y'. In this way, a concrete situation is converted into abstract variables and relations and intuitive reasoning is transformed into formal reasoning. For example, in everyday life one can say 'the greenhouse effect will be enhanced in the coming decades', but in a dynamic formal model the statement must be made specific by stating clearly defined variables and relations, for example 'the greater the amount of greenhouse gases in the atmosphere, the higher the temperature will be.'

At the level of the system, the behavior over time may be difficult to predict. The behavior of a dynamic system is determined by the constituent parts of the relations, but the behavior is not always a simple addition of effects. A system with even just a few variables can already show complex dynamic behavior. Step by step reasoning such as 'a higher temperature of the Earth leads to greater outgoing radiation which leads to lower energy of the Earth, leading to a decrease of the Earth's temperature', is clearly not enough to understand what the state of the system will be in the end. For example, the occurrence of an equilibrium temperature is hard to derive in this way. For that derivation, the relative contributions of the different effects must be taken into account, in order to predict if and how they will level out (Löhner, 2005). The computer simulation of the model can show the behavior of the system in the long term. Combining reasoning and predicting with interpreting the simulated data gives the learner insight into the different levels of the dynamic system (Wilensky & Resnick, 1999), from simple direct relations that can be reasoned with easily to complex relations that are composed of multiple steps that are more difficult to arrive at by reasoning with the constituent parts. When a complex composite model is difficult to reason with, a 'synthesizing' thinking strategy can be applied in which the model is viewed and evaluated as a whole (Stratford et al., 1998).

By observing the mechanisms that emerge from the composite structure of the model, underlying mechanisms may be discovered. (Wilensky & Resnick, 1999). Reasoning with single relations can be combined with reasoning at a composite level. For example, one can reason about the energy flowing out of the Earth's energy system and at the same time reason about the Earth's temperature rising. Research has shown that modelers of causal networks learn about complex systems by first learning the different fragments of a causal model and later integrating the different pieces in an interconnected complex causal model (Hagmayer & Waldmann, 2000). Mental simulation of the system plays an important role in learning about complex dynamic systems (Sterman, 1994). It appears that understanding the complex causal structures is supported by generating predictions based on mental simulation (Hagmayer & Waldmann, 2000).

An assessment that measures insight into dynamic systems has been developed by Booth Sweeney and Sterman (2000) in their systems thinking inventory. In their view thinking systemically means understanding systems concepts such as feedback, delays, and stocks and flows. In an empirical study in which highly educated students had to perform an 'extremely simple' task for 20 minutes, the learning of these basic systems thinking skills appeared to be a demanding task. The assessment was limited in time and did not require the creation of a model.

In summary, our exploration of the different aspects of computer modeling distinguishes three main components: scientific reasoning skills, conceptual domain knowledge, and insight into dynamic systems. In order to assess these components of the learning outcomes of computer modeling in a test, the assessment must meet several criteria: it must be valid and reliable; the items must be able to measure specific aspects within the components; and the items' scoring method must be reliable. In the next section, we discuss aspects of test theory that are relevant to these criteria.

1.3 The assessment of modeling-based learning outcomes

Answering the question of how to measure the specific learning outcomes of modeling is the core issue of this dissertation. Our goal is to develop, with the help of test theory, a standardized test, covering the wide spectrum of expected learning outcomes of modeling applied in a model-based context. In order to develop an 'ideal' test instrument, each scale of the instrument needs to be relevant, accurate, unbiased, sensitive, unidimensional, and efficient (Polit & Hungler, 1991). Though most instruments will not match this ideal, there are a number of techniques to evaluate the quality of the measurement instrument or, in other words, to *validate* the instrument. In Chapter 2 we will present the test as we developed it. In this section we will describe the evaluation techniques that were relevant in our research.

1.3.1 Validity

An important criterion in evaluating the quality of a test instrument is *validity*. In test theory, validity has been described in many ways, all having to do with the degree in which a test is a correct operationalization of the intended construct. Whether the test results can be interpreted as representing the construct is called the validity of the test (Stouthard, 1998). A general definition is posed by Messick (1989) who writes: 'Validity is the integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment' [emphasis in original] (p. 13). In our work we have assessed the validity of our test in two ways.

First, we used a “bottom-up approach” by making a detailed analysis of psychometric characteristics of our items in relation to the theoretical aspects of model-based reasoning outcomes that we distinguished. We used item response theory for this, which helped us to find evidence of the structure of the theoretical framework. As item response models are theory driven, developing a valid test based on item response theory starts with a clear definition of the constructs to be measured (Liu & Boone, 2006). The constructs can then be tested by fitting the test data to the item response model.

Second, we investigated the discriminative power of the test with respect to groups of students who are expected to perform differently on the test based on their personal characteristics or on experimental conditions they are in. This is also called the *known-groups technique*, an approach to analyze construct validity (Polit & Hungler, 1991). This technique administers the test to groups who are expected to perform differently on the test.

The many different forms of validity indicate the versatility of the concept of validity and the fact that it is not an 'all-or-nothing' characteristic of an instrument. Validity is a question of degree: it is not 'proved' or 'established', but instead supported by several forms of evidence (Polit & Hungler, 1991).

1.3.2 Reliability

Besides validity, the quality of a measurement instrument is determined by the *reliability* of the instrument. Reliability can be defined by the degree of consistency in the test outcomes (Polit & Hungler, 1991).

An initial requirement for a consistent test is a well-defined scoring procedure that classifies the test responses into a set of standard scoring categories. A scoring mechanism with sufficient detail is required to facilitate consistent scoring by different raters and improves interrater reliability or reproducibility. There are a number of interrater reliability measures, for example Cohen's kappa (Cohen, 1960) and Krippendorff’s alpha (Hayes & Krippendorff, 2007).

One way to analyze the degree of consistency of test scores is to analyze the internal consistency. The most common measure for estimating internal consistency is the coefficient Cronbach's alpha (Cronbach, 1951). Generally, alpha will increase when the inter-item correlations increase. Furthermore, Cronbach's alpha is a function of the number of items. In our case, where we are developing a composite test based on a framework with many aspects, the reliability of each of the constituent parts is important.

Similar to the concept of validity, reliability is not a fixed characteristic of a test. Reliability might be investigated for a specific set of response data collected in a specific context and under specific conditions.

1.4 Purpose of our research

In the previous sections, we have argued that, in spite of its many acclaimed advantages, there is little rigorous evidence about the specific learning outcomes of modeling. This lack of evidence forms a serious impediment to the decision to accept and further implement modeling approaches in secondary education. Therefore, in this dissertation we will investigate the specific learning outcomes of computer modeling in the context of a high school modeling activity. We develop a framework to describe and measure the specific learning outcomes. As noted above, our main research question is: What specific learning outcomes can be expected from computer modeling and how can they be measured?

In Chapter 2 the expected learning outcomes of modeling are defined in detail. We introduce the ACE framework for modeling knowledge, describe its components, and present a test that is based on this framework, called the ACE test.

In Chapter 3, we report on a validation study that investigates the validity of the ACE test. We analyzed the response data of students with different levels of modeling proficiency to find evidence for the validity of the test. Furthermore, we investigated the power of the test to discriminate on the main aspects of the framework between groups of students with different backgrounds in modeling.

In two comparative studies, the ACE test was used to investigate the differences and similarities between different modes of instruction. We chose modes of instruction that were progressively more closely related and thereby provided an increasingly rigorous evaluation of the discriminative power of the test.

In Chapter 4, we describe the first comparative study, in which we investigate the discriminative power of the test with respect to the comparison of expository instruction with modeling-based instruction. We expected to see differences in performance on the typical modeling aspects of the test concerning the construction of models and experimentation on the one hand, and the more 'traditional' aspect of reproducing knowledge on the other hand.

In the second comparative study we further investigated the discriminative power of the test with respect to the comparison of two modes of instruction that are more similar than the two modes of instruction in the first comparative study. In Chapter 5, this second comparative study is described in which we compared modeling with simulation-based instruction. These conditions differ only on the aspect of learners' construction of models, and are similar with respect to exploration and evaluation of a model. As in the first comparative study, we expected the construction of a model to lead to constructive skills. Furthermore, we expected no differences on test items related to performing experiments and evaluating data, because experimentation is involved in both modes of learning.

Chapter 6 presents an additional combined analysis of the data of the two comparative studies. This allows comparison of the modes of instructions over both studies.

In Chapter 7 the results and conclusions of the separate studies are recapped and discussed. The combined analysis is reviewed and implications for future research are presented.

Chapter 2 A framework for the assessment of learning by modeling[1]

Abstract

Learning by computer modeling is claimed to yield learning gains in the fields of knowledge of dynamic systems, higher-order reasoning, and domain-specific knowledge. However, it is hard to substantiate these claims with objective test measures. It is our aim to develop a test for these knowledge types that is able to detect differential learning outcomes between different modes of instruction (e.g. modeling, learning with simulation, and expository teaching). We present a framework to distinguish learning outcomes on three dimensions: type of reasoning process, complexity of knowledge elements, and domain-specificity. Based on this framework, we propose a test with specific items for each of the resulting combinations.

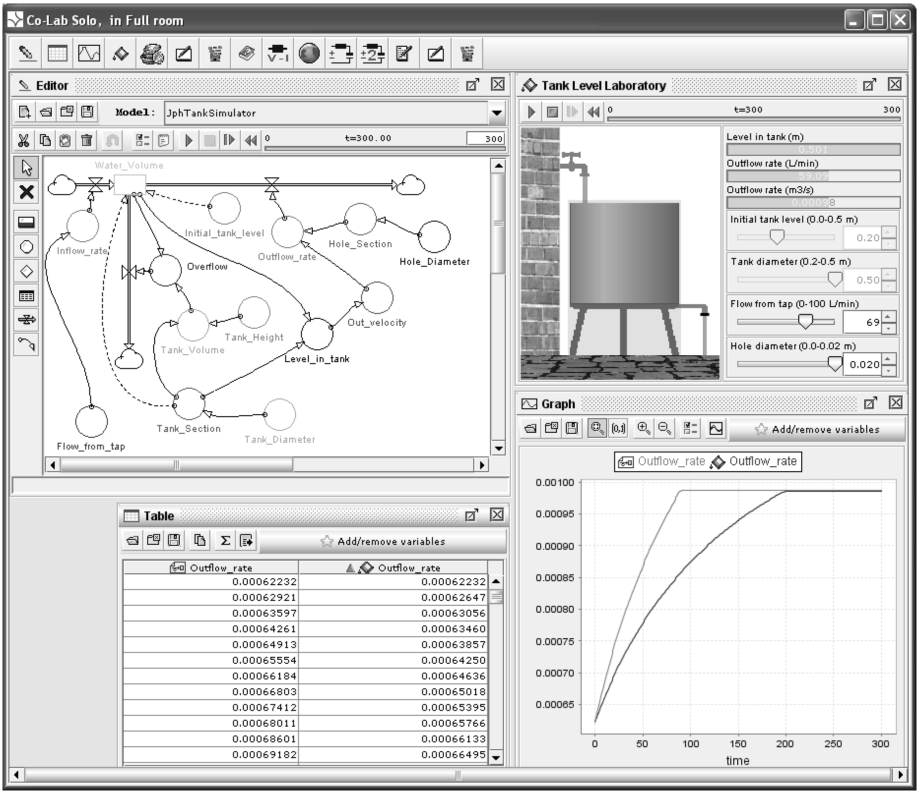
[1] This chapter is adapted from Van Borkulo, S. P., Van Joolingen, W. R., Savelsbergh, E. R., & De Jong, T. (2008). A framework for the assessment of learning by modeling. In P. Blumschein, J. Stroebel, W. Hung & D. Jonassen (Eds.), *Model-based approaches to learning*. Rotterdam, Netherlands: Sense Publishers.

2.1 Introduction

Current trends in secondary science education stress the importance of knowledge construction by students through active interaction with the learning environment (Blake & Scanlon, 2007; Gomez, 2005; Quintana et al., 2004; Tobin & Tippins, 1993). In science education, scientific reasoning and critical thinking play an important role as part of such knowledge construction (Fretz et al., 2002; Jacobson & Wilensky, 2006). This has led to an interest in *computer modeling of dynamic phenomena*, in which one acquires an understanding of the domain at hand by building, using, and testing computer models (De Jong & Van Joolingen, 2007; Löhner, 2005; Sabelli, 2006). Dynamic phenomena are defined as phenomena that reveal autonomous behavior as a function of time and often feature interactions between variables, feedback loops, nonlinear behavior, and time delays. The behavior of such phenomena is hard to predict or understand by reasoning alone (Hmelo-Silver & Azevedo, 2006; Kuhn, 2007). In the natural sciences this has led to the emergence of the field of *computational science*, in which the understanding of complex systems is based on building and simulating computer models. This has been applied to many kinds of complex systems, such as the weather system, global warming, molecular dynamics, and population biology. Computer simulation and modeling gradually became more than just a useful tool, as it opened up new fields of research and new methodological approaches in many fields of science. As the chief of the British Natural Environment Research Council put it, we need "a 'new breed' of scientist, and new ways of problem solving that cut across traditional disciplines" (Masood, 1998).

It would be desirable to have some of these developments reflected in the classroom. Computational science in secondary education became feasible through the introduction of computer modeling tools such as STELLA (Steed, 1992), ModelIt (Jackson, Stratford, Krajcik, & Soloway, 1996) and the Co-Lab modeling tool (van Joolingen, de Jong, Lazonder, Savelsbergh, & Manlove, 2005) (see Figure 2-1). Using such tools students create models to represent their ideas about a domain, in terms of variables and relations. The model can be run, to compute the values of the variables as they develop over time. This allows the modeler to see the course of events predicted by the model. The modeler evaluates the calculated values by comparing them to his or her predictions or to data collected in an experiment, and consequently, may accept the model, or decide to revise the model and start the cycle over again.

Students' modeling activities, as they bear close parallels to 'real' science, can become an integrated part of science education, and will partly define the learners' scientific ways of thinking. The focus of the current chapter is the occurrence of modelers' scientific knowledge and reasoning under the influence of a modeling task. In other words, what is learnt from computer modeling. In an overview of modeling research, Löhner (2005) distinguishes several expected effects of modeling on learning outcomes. These include a better understanding of the behavior of dynamic systems in general (Forbus, 1996; Hogan & Thomas, 2001; Stratford et al., 1998; Wilensky & Resnick, 1999), the development of specific (scientific) reasoning skills (Milrad et al., 2003), and domain-specific knowledge.



Model: JphTankSimulator

Level in tank (m)

Outflow rate (L/min)

Outflow rate (m3/s)

Initial tank level (0.0-0.5 m) 0.20

Tank diameter (0.2-0.5 m) 0.50

Flow from tap (0-100 L/min) 69

Hole diameter (0.0-0.02 m) 0.020

Outflow_rate	Outflow_rate
0.00062232	0.00062232
0.00062921	0.00062647
0.00063597	0.00063056
0.00064261	0.00063460
0.00064913	0.00063857
0.00065554	0.00064250
0.00066184	0.00064636
0.00066803	0.00065018
0.00067412	0.00065395
0.00068011	0.00065766
0.00068601	0.00066133
0.00069182	0.00066495

Figure 2-1 The learning environment of Co-Lab with its modeling tool

A number of studies have investigated the processes of modeling as a learning activity. For instance, Löhn er, Van Joolingen, Savelsbergh, and Van Hout-Wolters (2005) and Sins, Savelsbergh, and van Joolingen (2005) studied students' reasoning during the modeling process by analyzing students' conversations during a collaborative task; Hogan and Thomas (2001) focused on students' cognitive behavior while modeling, for example students' focusing on the output of a model or deciding to use either variables or constants. Less focus has been directed to the specific *learning outcomes* of modeling as a learning activity (Löhner, 2005). Spector, Christensen, Sioutine, and McCormack (2001) found that most researchers think the standard measures of learning are not adequate for a serious evaluation of learning in these learner-centered modeling environments for complex domains. This calls for the design of more adequate instruments that are able to measure the specific types of knowledge that are built by learners as a result of a modeling activity. An appropriate test for knowledge and skills related to modeling will not only contribute to a better research in the subject of modeling, but is also necessary for a full acceptation of modeling in educational practice.

In order to develop such an instrument at secondary education level, we need a precise operationalization of the knowledge and skills involved in utilizing models for understanding scientific phenomena. Such an operationalization provides a basis for designing instruments to detect the development of specific modeling knowledge. These considerations result in two questions:

What specific learning outcomes can be expected from computer modeling of dynamic systems?

How can these learning outcomes be measured?

Based on an analysis of the processes of computer modeling and reasoning with models we present a framework that serves as a model for modeling knowledge.

2.2 Modeling modeling knowledge

In describing modeling knowledge, we focus on the reasoning activities taking place in the modeling process (Löhner et al., 2005; Sins et al., 2005). Our basic assumption is that the core of modeling knowledge is (1) the ability to use models in reasoning about scientific phenomena and (2) the ability to perform the relevant reasoning processes that lead to new models. Both aspects of modeling knowledge include the model construct as a source of student behavior and can be considered to be essential for the development of modeling proficiency.

In the current section we therefore investigate the reasoning processes involved with using models in scientific domains and describe the influence of complexity and the domain on these reasoning processes.

2.2.1 Dynamic systems and scientific reasoning

Dynamic systems are characterized by an autonomous change over time of the system’s state. The variables in the system determine the process underlying this autonomous change. A simple example of a dynamic system is a bucket, filled with water, with a hole in it (see Figure 2-2 and Figure 2-3). Water will flow out through the hole and the system's state, represented by the water level, will change over time. The diameter of the hole is a parameter that determines the rate of change of the water level. More complex examples, requiring more variables to represent the state and to describe the change of the system are, for instance, the weather above Europe, the population of foxes and rabbits in a specific area, or the air conditioning system in an office building.

Reasoning with dynamic systems is centered around the characteristics of models of dynamic systems and is also called *model-based reasoning* (Magnani, Nersessian, & Thagard, 1998). The models we are discussing here are theoretical conceptualizations consisting of variables and relations that describe the behavior of a phenomenon. Reasoning with these models means to construct arguments using these variables and relations, for example explaining the decrease of the water level in the water bucket by naming the outflow variable or, in a more complex situation, referring to a feedback loop to argue the presence of equilibrium.

In the theories of scientific reasoning of Klahr and Dunbar (1988) and van Joolingen and de Jong (1997), scientific reasoning is seen as searching a hypothesis space and an experiment space. The hypothesis space represents all possible hypotheses about the domain, whereas the experiment space represents all experiments that can be performed in the domain. In this description a model would be an element of hypothesis space with the relations in the model and the characteristics of the dynamic system representing hypotheses, whereas the data evaluating the model, would originate from experiment space.

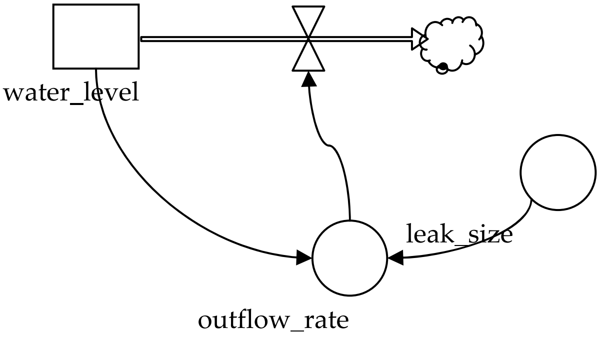
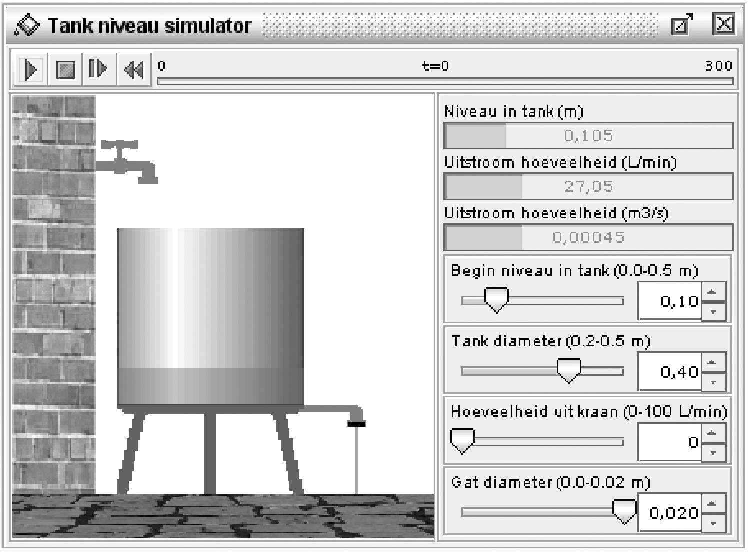


Figure 2-2 A system dynamics model of a leaking water bucket



Niveau in tank (m) 0,105

Uitstroom hoeveelheid (L/min) 27,05

Uitstroom hoeveelheid (m3/s) 0,00045

Begin niveau in tank (0.0-0.5 m) 0,10

Tank diameter (0.2-0.5 m) 0,40

Hoeveelheid uit kraan (0-100 L/min) 0

Gat diameter (0.0-0.02 m) 0,020

Figure 2-3 Simulation of a leaking water bucket

In reasoning with models it is essential to be able to move within and between these spaces. In using the model to predict or explain behavior of the investigated phenomenon, one makes a relation between the two spaces, linking the model to experimental outcomes. Moves within the hypothesis space represent changes in the model. Moves within the experiment space represent the search for empirical evidence for a model.

In the following we will argue that there are three basic reasoning processes in modeling: *Applying* a model, *Creating* a model, and *Evaluating* a model. Hence we will refer to our framework as the ACE framework. The basic processes are elaborated below.

2.2.1.1 Apply

Applying a model refers to using the model to generate outcome. The dynamic nature of the models in our focus implies that time-dependent behavior is an essential element in reasoning. With respect to time-dependent behavior we distinguish *prediction* and *explanation* as core reasoning processes (Löhner et al., 2005; Sins et al., 2005). Predicting means to infer the future behavior of the system from a given state, yielding propositions as: "the temperature will increase", or "the water level will reach equilibrium". This also includes predicting changes in behavior as a consequence of changes in parameters. Predictions can be supported by arguments, such as "… because the inflow will reach the same value as the outflow". Explaining can best be described as 'post-dicting', i.e. supporting the observed behavior by properties of the model, such as: "The water level has increased because the inflow has increased. This leads to rising water level which increases the outflow. When the outflow equals the inflow equilibrium is reached." Both predicting and explaining are based on a given model, classifying them as kinds of *applying* a model to a given situation. Predicting generates experimental data, and can be supported by using a simulation of the computer model. Explanation links the model to experimental data by focusing on the causes of observed behavior. Applying thus represents the relation between experiment and hypothesis space.

In applying knowledge of a dynamic system one must use the characteristics of a model to (*mentally*) *simulate* it. Mental simulation is executed by step by step thinking through causes and their effects. Using such mental simulation, applying models can lead to predictions of future model (or system) behavior, as well as explanations of observed behavior in reality. For prediction, the focus is on the result of the mental simulation (in terms of a description of the system's state and/or its development over time). For explanation, focus is on under what circumstances a certain behavior occurs. In the water bucket example, a prediction using a mental simulation could be: "Suppose the inflow from the tap increases, this would increase the water level, which will in its turn increase the outflow, which will decrease the water level, which will decrease the outflow, ... etc. Equilibrium will occur when the outflow again is equal to the new inflow. This equilibrium will be higher than its current value, because the outflow is positively related to the water level". An explanation of a water inflow increase could use the same mental simulation, but now *after* the fact. Therefore the reasoning is almost the same, but the focus of an explanation is not on the outcome but on the possible causes of an observed change in systems' state. This means explanation and prediction can be seen as two ends of the same process, with mental simulation of the model as the main reasoning mechanism.

2.2.1.2 Create

Predicting and explaining require a model being present. When this is not the case or when a given model is not adequate a model needs to be *created*, either from scratch or by modifying an existing model. This means searching hypothesis space for relations and variables that can represent the relevant parts of the phenomena being modeled. For instance when a system of a leaking bucket is extended with an extra bucket that collects the water flowing out of the first bucket, modelers need to add a second state variable representing the level in the new bucket and a relation connecting it to the first one. In doing so, besides phenomenon-related knowledge (for example knowledge of fluid dynamics), general conceptual modeling knowledge is involved. One needs to know how to define a new variable, how to use the different types of variables, and how to relate all components into an effective model.

2.2.1.3 Evaluate

Finally, by searching experiment space, modelers can *evaluate* a model by checking predictions that are generated based on the model against the actual behavior of the system that is being modeled. When students evaluate a model, they engage in *scientific reasoning* (Löhner et al., 2005). This basically means testing the hypothesis that (part of) the model is adequately describing the system that is modeled, considering prior knowledge of the domain or by comparing to reality or a simulation of reality. This involves designing and performing experiments, evaluating the results, and drawing conclusions, with respect to the model that has been created. For example, given the situation of a water bucket represented by a model with the state variable 'water level', and a flow which is specified by the relation: 'the higher the water level in the bucket, the faster the water will flow out', the student should be able to set values for the water level to test this hypothesis. This means, that the student should be able to observe the results and draw a conclusion on the correctness of the hypothesis.

The three ACE processes described here define the basic reasoning processes of modeling. The performance in these processes can be moderated by characteristics of the system that is being modeled. We consider two of these factors: the complexity of the modeled system and domain-specific knowledge.

2.2.2 Complexity of Dynamic Systems

Models are especially suitable for the managing and understanding of systems that have a certain complexity (Forrester, 1994). When the number of variables and relations becomes larger, a point will be reached in which it becomes impossible to infer system's behavior by reasoning alone. New behavior may emerge from the combinations of effects of individual relations. For instance, in the example of the leaking water bucket, the process of outflow is the only process in the system, whereas in an example of population dynamics, processes of birth, predation, and natural death may interact with each other such that either the population reaches equilibrium or an oscillatory system emerges. Dynamic modeling tools allow for individual specification of parts of the complex system, e.g. in terms of individual relations and variables. Simulation of these models makes clear what the effects are of all processes taken together on the system's state.

The complexity of a model depends on the number of relations that determine a variable (e.g. in a population model birth rate may depend on many factors), the 'length of the chain' of intermediate variables in a step by step reasoning, the presence of feedback, and the presence of counteracting effects. Indirect relations are complex and computationally demanding (Glymour & Cooper, 1999). Moreover, in a complex structure it may be difficult to derive the behavior of the complete system from the behavior of the individual relations. In systems consisting of many interconnected subsystems, one may need to reason in terms of global behavior that is not explicitly represented in a model, e.g. equilibrium (Wilensky & Resnick, 1999).

It is clear that for complex systems *mental* simulation becomes impossible. However, computer simulation can help understanding such systems. As emergent behavior (Holland, 2000) is displayed, the modeler can start to reason in a more abstract way about the system. This abstraction entails describing system behavior in terms of behavioral and time-independent terms such as "equilibrium", "oscillation", and properties that qualify these terms, such as "equilibrium value", "amplitude". This means that the level of description changes from pure causal reasoning for "simple" systems ("if A increases, then B decreases") to more holistic descriptions like the ones presented.

The consequence is that the reasoning processes that we introduced in the previous section change as a result of implications of complexity. The mental simulation of the model will be more difficult, the reasoning more abstract, and global behavior needs to be taken into account. Therefore, whenever assessing learning outcomes of modeling, complexity needs to be taken into account.

2.2.3 Domain-specific knowledge

Model-based reasoning to a high extent depends on domain-specific knowledge. Although in principle the behavior of the model can be studied at an abstract level (e.g., a model's behavior does not change when we change the names of the variables) and reasoning can take place at the abstract level as well, domain-specific knowledge will interact with the reasoning as soon as the model has a domain-specific interpretation (Fiddick, Cosmides, & Tooby, 2000; Westbrook, 2006). In fact, content-specific reasoning relies on links between declarative knowledge about the domain and procedural knowledge, in this case about modeling (Evans, 1989).

As an example, consider the leaking bucket model. An isomorphic model can be made of a discharging capacitor (with Q, the charge on the capacitor replacing the water level and 1/R, the inverted resistance in the discharging circuit being the 'leak size'). People not familiar with electronics will probably display different reasoning behavior for this model than for the bucket model, as for the bucket model they will be able to bring up a mental representation of the modeled system.

Prior domain knowledge influences the reasoning processes. Experts have 'chunked knowledge' (Chase & Simon, 1973), which means that larger blocks of knowledge form a whole. This enables experts to quickly establish a relation between the presented context and their mental models (Chi, Feltovich, & Glaser, 1981). The chunks enable a more efficient reasoning, which means that reasoning steps will be different, steps in between will be skipped, and shortcuts in reasoning will be taken.

Domain knowledge could also lead to disadvantages, such as biases induced by previous observations and beliefs about the domain. Moreover, prior knowledge may contain naïve conceptions or preconceptions, and alternative conceptions or misconceptions (Chi & Roscoe, 2002).

As a consequence, general and domain-specific knowledge are hard to disentangle. Reasoning with a model for which domain knowledge is available will partly invoke general reasoning skills and partly be based on stored domain-specific knowledge, which may facilitate or contaminate the modeling process. The corollary to this is that assessment of general modeling skills should take place outside the context of known domains.

2.3 The ACE Framework

Based on the preceding observations, we developed a framework to describe modeling knowledge more formally. The purpose of the framework is to be able to structure the types of knowledge that can be gained from learning by modeling. The framework can then be used to construct tests or observation schemes to assess students' modeling performance. As such we hope the framework will help the field in studying the development of knowledge under varying conditions of modeling.

The framework will be built on three dimensions: 1) reasoning process, 2) complexity, and 3) domain-specificity. Each will be briefly described below.

The first dimension, reasoning processes, entails the extent to which someone can perform the reasoning processes related to modeling: apply, create, and evaluate models. This dimension covers the basic reasoning processes of modeling as was described by the ACE model: applying knowledge of a model (predicting and explaining), creating (or modifying) new models, and evaluating models and data generated by models.

The second dimension, complexity, concerns the complexity of the systems to which the knowledge element applies. A model has complexity due to its behavior and due to its structure and this influences the reasoning processes. We found behavioral, 'holistic' aspects, for instance the occurrence of equilibrium and oscillation, and structural aspects, for instance the number of variables and relations, the number of steps in a step by step reasoning. To cope with these aspects of complexity, we define *simple* as the smallest meaningful unit of a model, with one dependent variable and direct relations to that variable only (see Figure 2-4), and *complex* as a larger chunk that contains indirect relations and maybe (multiple) loops (see Figure 2-5). An important aspect in complexity is the length of the paths that can be followed in the reasoning with a model.

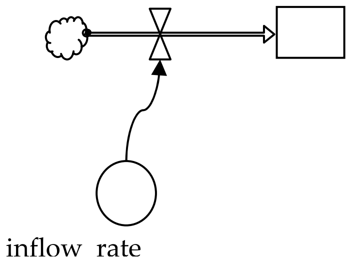


Figure 2-4 Example of a simple model

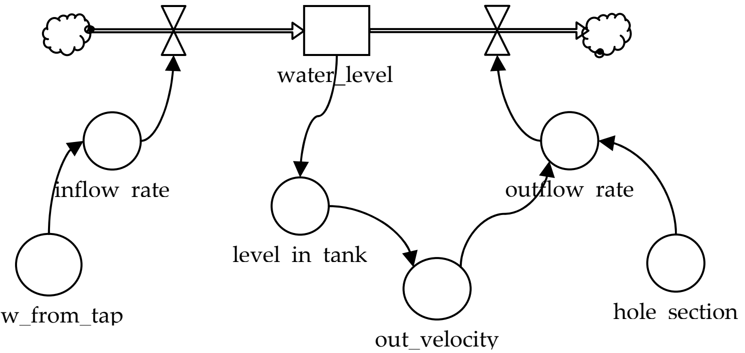


Figure 2-5 Example of a complex model

Finally, the third dimension, domain-specificity, relates to whether a knowledge element is specific for a domain or has a more general, transferable nature. Domain-specific and general reasoning skills are interwoven and function in reciprocal interaction. Therefore, general reasoning skills might best be articulated in a domain without prior domain-specific knowledge, for example in a domain about a fantasy topic.

Combining the three dimensions now yields the complete 'ACE framework' with twelve cells: each of the three types of reasoning can be applied to simple and complex models, and can be related to a specific domain or be of a general nature.

2.4 Applying the framework

In this section, we will apply the framework to an introductory model in the domain of global warming, in order to illustrate how the framework can be used to describe the relevant modeling-related knowledge in this particular domain and how it can be used as a guideline to construct test items directed at measuring specific aspects of this knowledge.

2.4.1 Domain: global warming

We chose a topic that is part of the curriculum of pre-university students and in which it is possible to create a simplified model: global warming. In the basic energy model of the earth, a central concept is the *radiation of the sun*, providing *energy to the earth*, the earth losing energy because of the *outgoing radiation*. Other influencing variables are the degree of reflection of the surface of the earth, represented by the variable *albedo*, the *temperature* of the earth surface, and the *heat capacity* of the earth.

The underlying model should not be too simple or too complex. In our basic energy model there is one state variable, *the energy in the earth*, being the central variable. There are 6 auxiliaries: radiation of the sun, albedo, inflow of energy, temperature, heat capacity, and outgoing radiation (see Figure 2-6). The model shows a longest path length of 5, an inflow and outflow part, and one negative feedback loop in the outflow. This gives opportunities to develop both simple and complex items using direct and indirect relations, and using the feedback loop. The feedback loop represents the presence of equilibrium. In the system of global warming the temperature will be in balance with specific values of the variables in the long term.

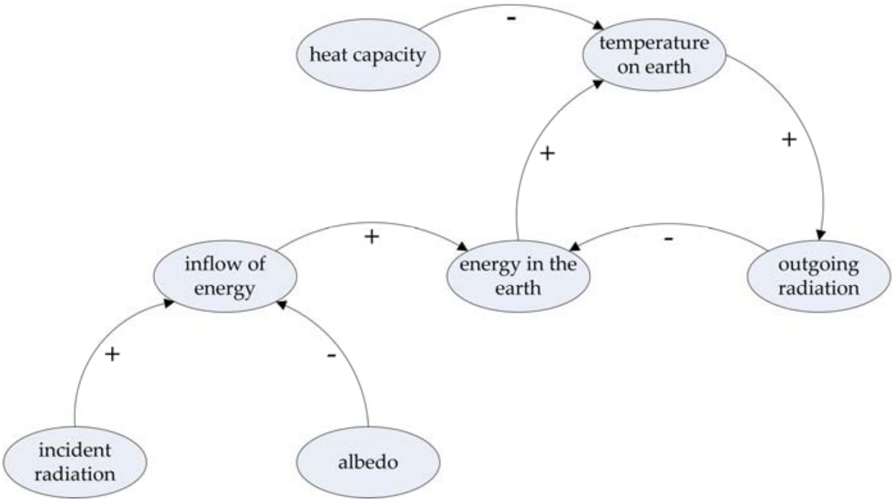


Figure 2-6 A basic global warming model

2.4.2 From framework to test items

We will describe items for the three types of reasoning: apply, create, and evaluate, and in two types of complexity: simple and complex. For this domain, global warming, the difference between simple and complex is made by using direct and indirect relations in the items. Furthermore, complex items include reasoning with equilibrium. To assess domain-specific knowledge, additional items on the reproduction of domain-specific knowledge were constructed.

2.4.2.1 Apply - simple

In the category 'apply - simple', the ability to reason with a direct relation is addressed. The sample item below assesses the understanding of the causal relation between albedo and the inflow of energy to the earth (see Figure 2-7). The relation between the two variables is represented in the given model as a direct relation. In addition to the multiple choice question, an explanation is requested to assess the understanding in more detail.

```
Choose the correct statement.
    A.  The higher the albedo, the larger the inflow of energy to the
        earth.
    B.  The higher the albedo, the smaller the inflow of energy to the
        earth.
    C.  The albedo does not influence the inflow of energy to the
        earth.
Explain your answer.
```

Figure 2-7 Example of an apply - simple item

Other possible operationalizations in the category 'apply - simple' are: applying a formula to deduce the value of a variable, or an open question that asks the student to explain a direct relation between two variables.

2.4.2.2 *Apply - complex*

In the category 'apply - complex', the ability to reason with indirect and composite relations is addressed. In the basic energy model of the earth the feedback loop is causing the occurrence of equilibrium. Similar to the 'simple' segment of the apply category, the ability to reason with complex structures can be assessed by a multiple choice item or an open question. The item below shows an example of an open question about the development of temperature in the given basic energy model (see Figure 2-8). The item asks the student to express understanding of equilibrium in both a sketch and an explanation.

We performed an experiment with a computer simulation. The values in the table show that the temperature on earth increased to 30 °C after three years.

Starting temperature	10	°C
Heat capacity	50	J/K
Inflowing radiation	3	W
Albedo	20	%
End values (after 3 years):		
Temperature	30	°C
Outgoing radiation	2.4	W

In a graph, sketch the development of the temperature between start time and end time of the computer simulation. Explain your sketch. Also, describe what happens with the temperature after a longer period of time.

Figure 2-8 Example of an apply - complex item

2.4.2.3 *Evaluate - simple*

The evaluation of models and data is addressed in the 'evaluate' category. In the 'simple' segment, items ask to evaluate a direct relation in a model or to evaluate specific data from an experiment with respect to a direct relation. In the sample item below, the causal relation between inflow of energy and albedo is asked to be evaluated (see Figure 2-9).

Is it correct that a high inflow of energy causes a low albedo? Explain why or why not.

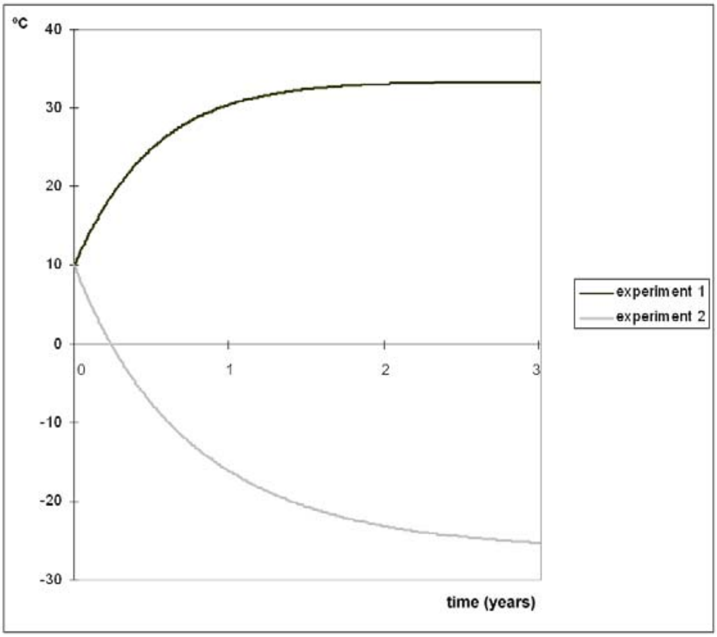
Figure 2-9 Example of an evaluate - simple item

2.4.2.4 Evaluate - complex

In the 'complex' segment, the evaluation of a larger part of a model or the evaluation of experimental data with respect to an indirect relation is addressed. In the sample item below, an experiment is described in which the indirect relation between the inflow of energy and temperature is investigated (see Figure 2-10). The item asks the student to evaluate a conclusion based on the presented experimental data .

André performed two experiments with the simulation to investigate the relation between inflowing radiation and temperature.

experiment 1:			experiment 2:		
Starting temperature	10	°C	Starting temperature	10	°C
Heat capacity	50	J/K	Heat capacity	50	J/K
Inflowing radiation	2.5	W	Inflowing radiation	1.5	W
Albedo	0	%	Albedo	30	%
End values (after 3 years):			End values (after 3 years):		
Temperature	33	°C	Temperature	-25	°C
Outgoing radiation	2.5	W	Outgoing radiation	1.1	W



From these data, André concludes that, the higher the inflowing radiation, the higher the temperature on earth. Is it correct for André to draw this conclusion? Explain your answer.

Figure 2-10 Example of an evaluate - complex item

2.4.2.5 Create - simple

In the 'create' category, the ability to create, modify or extend models is addressed. Creating at the simple level includes defining new variables, adding direct relations, and creating models consisting of only one dependent variable and direct relations with this variable. In the example below, the basic energy model is expanded with the concept of atmosphere (see Figure 2-11). The item asks the student to extract a relevant new variable from the context that models the presented situation.

> In our model, outgoing radiation is dependent on the temperature on earth. We did not take into account the fact that part of the outgoing radiation is not flowing out, but is reflected by the atmosphere around the earth.
>
> Describe which new variable(s) you need to model this situation.

Figure 2-11 Example of a create - simple item

2.4.2.6 Create - complex

Creating at the complex level includes creating larger model structures with indirect relations and feedback loops. In the sample item below, a situation is described of a watch running on solar energy with several functions and one of the functions, the display's lighting, giving information about the battery's energy (see Figure 2-12). The student has to draw a model of this situation. The purpose of the model is predicting the empty time of the battery.

> Suppose you have a watch running on solar cells. Just like a solar collector, the solar cells pick up the visible light of the sun, convert it to electricity, and store the electricity in the watch's battery. If you often use the watch's stopwatch function, the battery will empty earlier. Also, the intensity of the display's lighting is adjusted to the amount of energy in the battery. When the lighting is weak, you know the battery is almost empty.
>
> Draw a model of this situation in order to be able to predict the empty time of the battery. Explain your drawing.

Figure 2-12 Example of a create - complex item

The example items presented here demonstrate how all types of modeling knowledge distinguished in our framework can be operationalized based on a single introductory model in the domain of the energy of the earth.

For reliable use in an experimental study, each type of item needs a standardized scoring method. We designed a scoring method based on scoring correct and incorrect answer elements. Basis for the scoring method is the central element of a relation between two variables (Bravo, Van Joolingen, & De Jong, 2006; Löhner, Van Joolingen, & Savelsbergh, 2003). We divided the element of a relation into three parts: the existence, the direction, and the quality of the relation. Other answer elements are statements related to experimentation in the evaluate category items. Furthermore, answer elements strongly depend on the domain.

2.5 Discussion

In this chapter, we presented the ACE-framework for modeling knowledge, based on three dimensions: type of reasoning process (apply, create, and evaluate), complexity, and domain-specificity (general and specific). We demonstrated how the modeling knowledge framework enables the systematic design of an assessment instrument. For the domain of global warming, we applied the framework in the development of a test with items in the twelve cells of the framework. An important factor in knowledge building and mental models is prior domain-specific knowledge (Westbrook, 2006). Therefore, we added items to the test that accounted for conceptual knowledge in the domain of global warming.

By specifying the framework we categorized the learning outcomes that can be expected from the computer modeling of dynamic systems. This categorization supports the investigation of modeling knowledge in the different components of a modeling activity. The development of the test in the sample domain of global warming showed that the framework could be operationalized in a systematic way. The item construction in this domain gave a valuable and useful indication for systematic item development based on the framework.

Further research will need to focus on several aspects. First, the validity of tests based on the framework needs to be investigated. As a first step, we developed a test in the domain of global warming. The content covered by the test was checked by experts in the field of modeling to be representational within the domain. Hand in hand with the item development, we developed a systematic scoring method to ensure reliable scoring. In the scoring method, not only the correct elements are of interest, but also the occurrence of common mistakes and misconceptions within the domain. Further empirical studies need to focus on the psychometric validity and discriminative validity of the developed test.

Second, after having developed a test for one domain with promising results, the framework needs to be applied to other domains. Especially, the development of a test in a 'fantasy' domain in which students have no prior knowledge may give valuable additional information about the role of the domain dimension in the framework and give insight in the general reasoning skills related to modeling.

Chapter 3 Assessing students' model-based reasoning skills in the domain of dynamic systems: a validation of the ACE-test

Abstract

Computer modeling of dynamic systems is an emerging topic in science education. The assessment of computer modeling learning outcomes, however, still has some limitations to overcome. In this study, we validate a test about the specific reasoning skills in a pre-university level modeling task. The test is based on a modeling knowledge framework to distinguish between different types of reasoning (Apply, Create, and Evaluate), at different levels of complexity, and at different levels of domain-specificity. Test data of students with different levels of modeling experience were analyzed using a standardized scoring method with an acceptable interrater reliability. Construct validity was examined for the three types of reasoning (Apply, Create, and Evaluate) and additionally for Reproducing conceptual knowledge, using item response models. The analysis yielded evidence of the four unidimensional item scales. Significant differences in test scores were found as an effect of modeling experience and trends were found for domain knowledge. The results of this study suggest that the framework is suitable for developing a multifaceted test of modeling related knowledge and skills with discriminative power with respect to modeling experience. The ACE test appears to be suitable to distinguish between the specific model-based reasoning skills. Modeling reasoning abilities seem to relate to domain knowledge and might not simply be learned hand in hand with general academic reasoning abilities.

3.1 Introduction

Knowledge construction by students through active interaction with the learning material is gaining attention in current views on science education (Baggott La Velle, McFarlane, & Brawn, 2003; Kruckeberg, 2006; Park, 2008). Scientific reasoning and critical thinking play an important role as part of such knowledge construction and this has led to an interest in *computer modeling of dynamic phenomena*. In modeling, learners acquire an understanding of the domain at hand by building, using, and testing computer models (cf., Forbus, 1996; Löhner et al., 2003). Computer modeling allows moving beyond reproducing and applying knowledge as a learning goal towards goals related to creating and evaluating knowledge in domains that involve more or less complex dynamic phenomena.

Dynamic phenomena reveal autonomous behavior as a function of time and often feature interactions between variables, feedback loops, nonlinear behavior, and time delays. The behavior of such phenomena is hard to predict or understand by reasoning alone (Gentner & Stevens, 1983). Building and simulating computer models has helped to gain insight into many kinds of complex dynamic systems, such as the weather system, global warming, molecular dynamics, and population biology. Given the broad usefulness and the widespread application of these systems it is worthwhile for a student to be knowledgeable about computer models and gain insight in their dynamic nature.

Computer modeling tools such as STELLA (Steed, 1992) or Model-It (Jackson et al., 1996) provide the opportunity to introduce *computational science* in the upper secondary education classroom. Using such tools students create models to represent their ideas about a domain, in terms of entities, variables, and relations. When executing the model, the tool computes the values of the variables as they develop over time. This allows the learner to see the course of events predicted by the model. The modeler evaluates the computed values by comparing them to his or her predictions and, consequently, the modeler may decide to modify the model and start the cycle over again. Thus, computer modeling can become a valuable part of science education, just as it has become part of ′real′science.

Several beneficial effects of computer modeling on learning outcomes have been predicted (Löhner, 2005). These include a better understanding of the behavior of dynamic systems in general (Forbus, 1996; Hogan & Thomas, 2001; Stratford et al., 1998; Wilensky & Resnick, 1999) as well as the development of scientific reasoning skills (Milrad et al., 2003). Also, modeling can help the acquisition of domain-specific knowledge by learners (Schecker, 1998).

However, despite all these expectations, few claims have been substantiated (Löhner, 2005). A possible explanation for this lack of evidence may be that standard measures of learning are not adequate to assess the specific learning outcomes that can be expected in a modeling environment (Spector, 2001). Traditional forms of assessment in science education focus on reproduction and knowledge application to solve well-defined problems, mainly about static and simple dynamic situations. However, the added value of computer modeling education can be found in learning activities requiring dynamic reasoning about more complex dynamic phenomena. In these activities one develops abilities in defining relevant variables and relations, and in evaluating the created model by judging the individual relations, the structure of the model, and the data that is produced by the model.

Earlier attempts have been made to assess the learning outcomes of modeling, but these attempts suffer from several limitations in relation to science education research. For example, Booth Sweeney and Sterman (2000) concentrated on one aspect of modeling, systems thinking skills, learned in an 'extremely simple task' that lasted only 20 minutes. When students are learning to model for a longer period of time, more aspects of modeling will come into play and at a more elaborate level. At the other extreme some studies concentrate on themes that go beyond computer modeling such as 'nature of science', epistemological knowledge, and self-regulation (Guttersrud, 2006; Hogan & Thomas, 2001). Although computer modeling activities can contribute to these aims, this goes beyond the specific learning outcomes we are looking for in this study. Another limitation in previous research is the efficiency of the modeling assessment. Some studies assess modeling proficiency by means of information-poor students' log files (Ergazaki et al., 2005; Forbus et al., 2005) or use labor-intensive and subjective assessment methods such as interviews and videos (Forbus et al., 2005). In both cases the assessment is not useful in larger scale research and educational practice, and it is worthwhile to find more efficient methods.

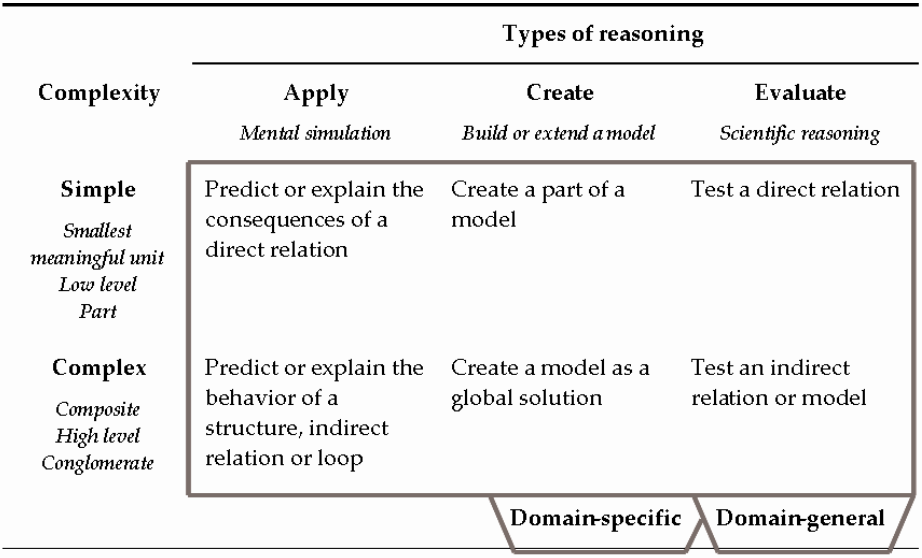
These limitations call for the design of more adequate instruments to measure the different types of knowledge and skills learners develop as a result of a modeling activity. An appropriate test for knowledge and skills related to modeling will not only contribute to better research in the subject of modeling, but is also necessary for a full acceptation of modeling in educational practice.

In order to be able to detect differences in learning outcomes between traditional and modeling education on these specific aspects, we have developed a test that incorporates both traditional learning merits such as reproducing, and modeling merits such as applying, creating, and evaluating. Based on the ACE framework we developed an assessment instrument with subscales for each type of learning outcome we distinguished (Van Borkulo, Van Joolingen, Savelsbergh, & De Jong, 2008). The current study aims to validate the instrument by using the instrument to assess the modeling knowledge of a diverse sample of students, who could be expected to differ in modeling proficiency.

3.2 The ACE Modeling Knowledge Framework

As follows from the description above, modeling knowledge involves the ability to perform specific reasoning processes in a domain, beyond reproducing domain facts and formulae and applying knowledge in the form of solving problems. This results in the ACE dimension, for *apply* (A), *create* (C), and *evaluate* (E) (Van Borkulo et al., 2008). Measuring these reasoning processes should reveal learners' knowledge about modeling and its underlying process.

Each of these ACE reasoning processes can be mastered in situations at different levels of complexity and in a domain-specific or a domain-general way. Both complexity and domain characteristics may modify the way the reasoning processes operate. Therefore these two aspects form a second and third dimension of the framework (see Figure 3-1).



	Types of reasoning		
Complexity	**Apply** *Mental simulation*	**Create** *Build or extend a model*	**Evaluate** *Scientific reasoning*
Simple *Smallest meaningful unit* *Low level* *Part*	Predict or explain the consequences of a direct relation	Create a part of a model	Test a direct relation
Complex *Composite* *High level* *Conglomerate*	Predict or explain the behavior of a structure, indirect relation or loop	Create a model as a global solution	Test an indirect relation or model

Figure 3-1 The ACE modeling knowledge framework with samples of modeling actions

3.2.1 First Dimension: Types of Reasoning (ACE)

The reasoning dimension describes the specific reasoning processes involved in the process of reasoning activities related to modeling. The relevant processes can be inferred from the theories of scientific inquiry (De Jong, 2006; Van Joolingen & De Jong, 1997; Klahr & Dunbar, 1988), where scientific reasoning is regarded as searching a hypothesis space and an experiment space. The hypothesis space represents all possible hypotheses about the domain, whereas the experiment space represents all experiments that can be performed in the domain. We distilled three types of reasoning, *apply* (A), *create* (C), and *evaluate* (E) models that constitute the core of modeling. First, *apply* refers to applying the rules of a model in order to infer predictions and explanations. In the case of dynamic systems, *applying* may involve the (mental) simulation of the evolution of the system over time. Both prediction and explanation are part of searching hypothesis space and relating the hypotheses found to experiment space. Second, *creating* a model is building a new model or expanding an existing model, by adding or modifying variables and relations in the model. Creation processes represent moving in and expanding the hypothesis space. Finally, *evaluating* a model represents a search in experiment space and the mapping of its results to the predictions made by the model.

Although our focus is on the role of these reasoning processes in dynamic systems modeling, their relevance goes beyond that, and in fact it should be noted that the reasoning processes involved in modeling bear strong resemblance to the higher order learning aims in Bloom's (revised) taxonomy (Anderson & Krathwohl, 2001).

3.2.2 Second Dimension: Complexity

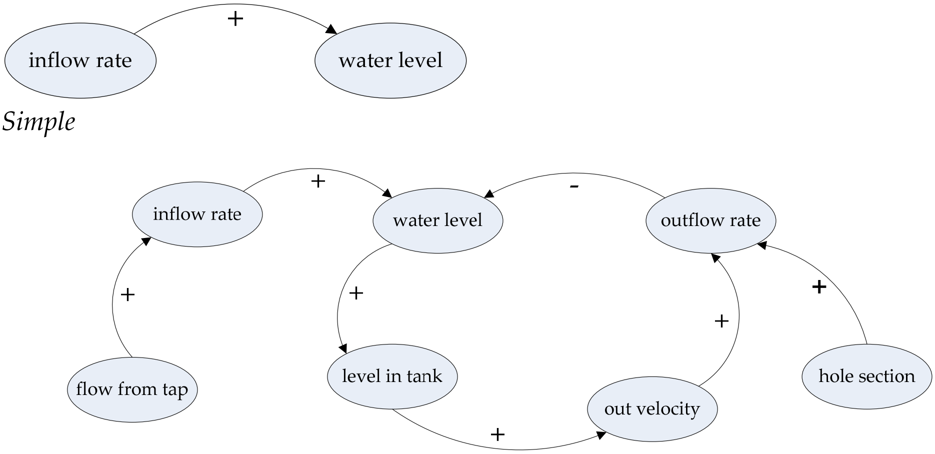
Each of the reasoning processes described above can be mastered either with simple models or also with complex models. Model complexity depends on structure and behavior. Structural aspects relate to the number of variables and relations, the number of steps in a step by step reasoning using the model, and the occurrence of circular structures, such as feedback loops (Gentner & Stevens, 1983; Johnson-Laird, 2001). Behavioral aspects include, for instance, the occurrence of equilibrium and oscillation.

Reasoning about a simple model structure is qualitatively different from reasoning about a more complex structure for each of the three ACE processes. For instance, applying knowledge of a simple model structure involves local reasoning such as "what are the direct consequences of a change in variable value using the relation 'the higher X, the lower Y'". When applying knowledge of a complex model structure, it is not sufficient to trace all direct relations. One also needs to take the overall structure into account, recognize the structure as a negative feedback loop, and weigh the parallel influence of several factors in the model. A prediction of the value of a variable (for example, 'variable X will approach a constant value') requires a step by step explanation of influences of factors (in the example, 'the higher X, the higher Y; but the higher Y, the lower X; and therefore the lower Y, etc...') and insight in the system's behavior (in the example, 'it is a feedback loop that causes equilibrium and therefore will reach a balanced value of X'). The qualitative difference between *simple* and *complex* reasoning is related to difficulty, and represents the fact that reasoning with complex structures requires higher-level thinking.

In the case of evaluation of a model or output data, the same distinction can be made between reasoning with simple and complex model structures. The evaluation of a simple model structure involves comparing experimental data produced by the model related to direct relations to reality or to one's own expectations and draw conclusions. The evaluation of a model with a more complex structure involves examining experimental data produced by parallel influences and system behavior or assessing a model structure.

In the case of creating a model, creating simple elements means defining new variables and direct relations between them. Creating complex structures involves more variables and relations to be added in a single coherent set of actions, for instance adding a substructure to a model that represents a complete feedback loop.

To cope with these aspects of complexity, we define *simple* as the smallest meaningful unit of a model, with one dependent variable and direct relations to that variable only, and *complex* as a larger chunk that contains indirect relations and possibly (multiple) loops and complex behavior, requiring more complex reasoning (see Figure 3-2).



Simple

Complex

Figure 3-2 Example of a simple and a complex model structure

3.2.3 Third Dimension: Domain-Specificity

Each of the reasoning processes can be mastered in a domain-dependent or in a domain-general way. In many cases reasoning is domain-dependent (e.g., Wason, 1968). Reasoning with models is no exception. In case of an abstract model, the student can only reason with the rules in the model, whereas in the case of a model framed in a concrete domain, the student's inferences will also be influenced by prior knowledge about the phenomenon (Sterman, 2002), for instance by retrieving known behavior of a system to support the generation of a prediction. Moreover, also at the strategic level, students' decisions and approaches will be steered by ideas and expectations derived from domain-specific knowledge (Chi & Roscoe, 2002). Domain-specific knowledge manifests itself in the ability to reproduce facts, formulae, and definitions that play a role in the domain.

3.3 Assessing modeling knowledge, the ACE test

In this section we describe the operationalization of the dimensions in the framework into test items. In order for the test to make a useful research instrument in secondary education settings, there are several requirements. First, the test should be easy to administrate, preferably without a need for specific software. Second, the test items must be described in a generic notation that can be understood by all learners without extensive training. Third, to give insight in reasoning used, the test items must also require an argumentation from the learner. Finally, in order to allow future usage in other domains, the item construction format must preferably be suitable for different domains and incorporate generic construction rules.

3.3.1 Types of Reasoning

The type of reasoning 'apply'. The ability to apply the rules of a model is expressed by translating the formal representation of the model into meaningful reasoning. This reasoning ability can be addressed by asking for explanations or predictions of model behavior, for example the influence of one variable on another (direct or indirect) or the occurrence of equilibrium. The explanation can be given by a step by step reasoning, using variables and relations in the model. Predictions can be given in words (e.g. "the temperature will increase", or in visual form, such as by drawing a graph representing the development of temperature over time.

The type of reasoning 'create'. The ability to create a model is composed of extracting relevant information from the context and translating the information into specific, computable variables and defining the relations between them. The context presented in a create item must use clear language in describing the context realistically, so that the student can show his or her ability to translate common language into the formal language of a model. The description of a realistic phenomenon usually contains more information than necessary for a formal model and is formulated less exactly. Therefore, create items need context descriptions that on the one hand give enough freedom in the translation from words to a model, but on the other hand are not too laborious in order to avoid too large an influence of linguistic competences.

The type of reasoning 'evaluate'. When evaluating a model or data, the focus must be put on a specific part of the model or specific data. For the specific part of the model that needs to be evaluated, the student is asked to argue whether the model adequately describes specific behavior. For example, 'Is it correct that a high value of X causes a low value of Y? Explain your answer.' When evaluating data produced by a model, the student needs to relate the data represented by a graph or a table to the representation of a *model* relation or hypothesis. These evaluating skills are evoked by questions such as 'Is it correct that conclusion Z follows from the presented data? Explain your answer.'

3.3.2 Complexity

In order to measure the effects of *complexity* for each reasoning type, items were created at different levels of complexity. Simple items involve direct relations with a limited amount of variables. Complex items refer to structures in the domain, involving multiple variables, indirect relations, and feedback loops. Complex items require reasoning beyond singular cause-effect relations and may involve global properties such as equilibrium.

3.3.3 Domain-specificity

In order to determine the effect of domain knowledge on reasoning, test items were created for both a real domain (in our case on the earth's greenhouse effect) and a fantasy domain (called harmony of the spheres). The models underlying both types of items are isomorphic. The fantasy domain is used to invoke reasoning processes on a model outside the context of available domain knowledge.

In order to control for the available domain knowledge it is worthwhile to assess the ability to reproduce conceptual knowledge (i.e. facts that can be known about the domain). Reproducing conceptual knowledge is evoked with questions such as 'What is concept X?' or 'What is the relation between concept X and Y? Describe the role of concept Z.' The possibility to construct reproduce items strongly depend on the concepts in the domain. Items on reproducing conceptual knowledge can only be created for the real domain.

3.3.4 Test Characteristics

The whole test was administrated in a paper-and-pencil format. As a notation to present the models in the test, we used the directed causal concept map (Novak, 1990). This format was understandable for all students after a short explanation. Variables are represented by circles labeled with a variable name, causal relations are represented by arrows, and the nature of the relation is expressed by a plus or minus sign. An advantage of using this format is that it is independent of specific modeling syntax used in the various modeling tools.

The entire test falls apart in a domain-specific part and a domain general part. Each part has an introduction in which the central model is introduced (see Figure 3-3), and is followed by the questions. The questions in the domain-specific part are about the greenhouse effect. In the model that precedes these question the Earth is treated as a "Black Sphere", which is the reason why this subtest was labeled "Black Sphere". This subtest includes 29 test-questions, including 6 questions about conceptual knowledge (examples are presented in Figure 3-4). The domain-general part concerning the fantasy domain "Harmony of the Spheres" is about the model presented in Figure 3-5, and consists of 23 questions (see figure 3-6 for examples). Table 3-1 presents the distribution of test items over types of reasoning processes.

Table 3-1 Distribution of the number of items for the black sphere domain and the harmony domain

Number of items	Black sphere domain		Harmony domain	
	simple	complex	simple	complex
Reproduce items	3	3	n.a.	n.a.
Apply items	3	4	3	4
Create items	3	6	3	6
Evaluate items	3	4	3	4
Total	12	17	9	14

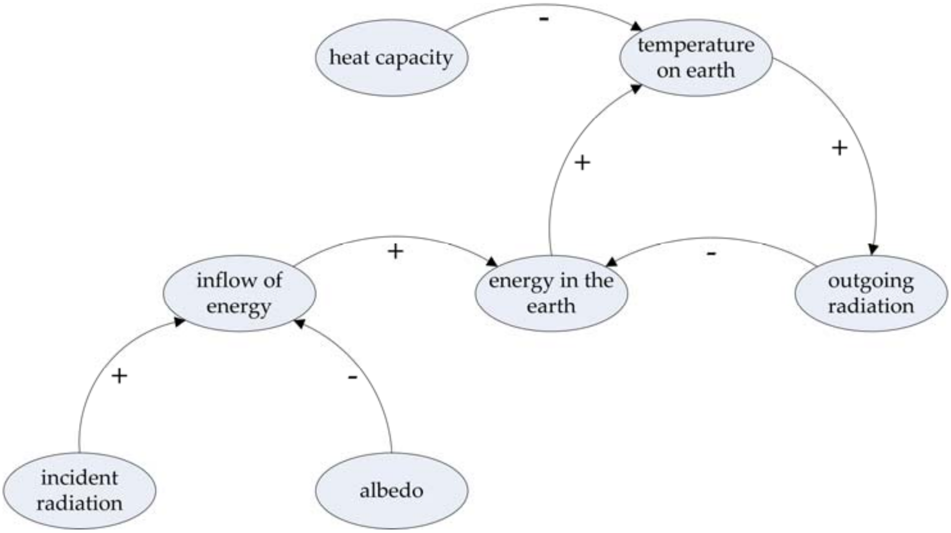


Figure 3-3 The introduction model of the black sphere that was given in the test

Reproduce - simple
What is albedo (or reflectivity) of a substance?

Reproduce - complex
What is the relation between energy in the earth and the temperature on earth? Describe the role of heat capacity.

Figure 3-4 Eight black sphere test items

Apply - simple

Choose the right statement.

A. The higher the albedo, the larger the inflow of energy to the earth.

B. The higher the albedo, the smaller the inflow of energy to the earth.

C. The albedo does not influence the inflow of energy to the earth.

Explain your answer.

Apply - complex

We performed an experiment with a computer simulation. The values in the table show that the temperature on earth increased to 30 °C after three years.

Starting temperature	10	°C
Heat capacity	50	J/K
Inflowing radiation	3	W
Albedo	20	%
End values (after 3 years):		
Temperature	30	°C
Outgoing radiation	2.4	W

In a graph, sketch the development of the temperature between start time and end time of the computer simulation. Explain your sketch. Also, describe what happens with the temperature after a longer period of time.

Create - simple

In our model, outgoing radiation is dependent on the temperature on earth. We did not take into account the fact that part of the outgoing radiation is not flowing out, but is reflected by the atmosphere around the earth.

Describe which new variable(s) you need to model this situation.

Create - complex

Suppose you have a watch running on solar cells. Just like a solar collector, the solar cells pick up the visible light of the sun, convert it to electricity, and store the electricity in the watch's battery. If you often use the watch's stopwatch function, the battery will empty earlier. Also, the intensity of the display's lighting is adjusted to the amount of energy in the battery. When the lighting is weak, you know the battery is almost empty.

Draw a model of this situation in order to be able to predict the empty time of the battery. Explain your drawing.

Figure 3-4 Continued

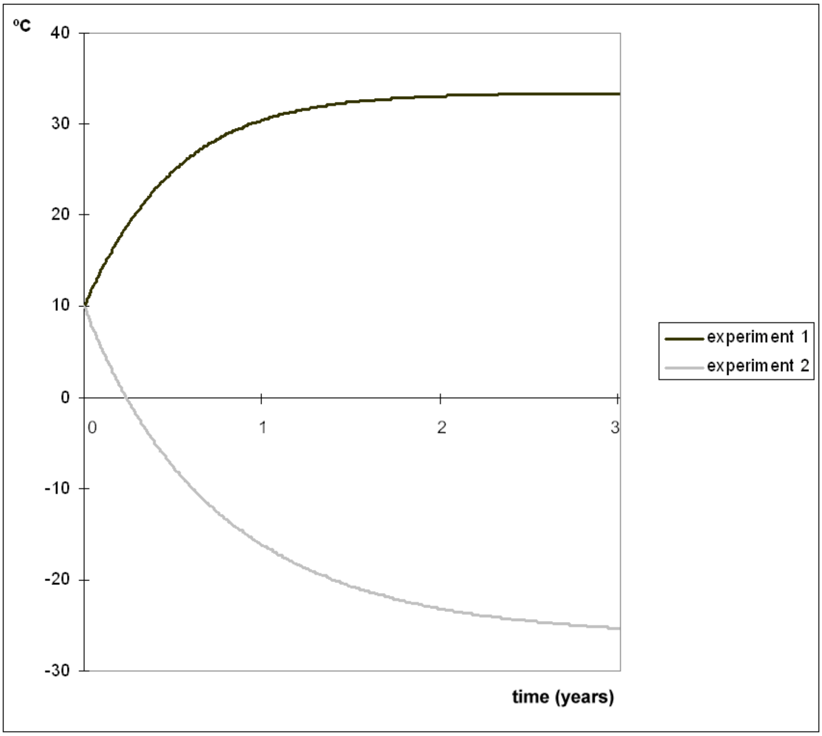
Evaluate - simple

Is it correct that a high inflow of energy causes a low albedo? Explain why or why not.

Evaluate - complex

André performed two experiments with the simulation to investigate the relation between inflowing radiation and temperature.

experiment 1:			experiment 2:		
Starting temperature	10	°C	Starting temperature	10	°C
Heat capacity	50	J/K	Heat capacity	50	J/K
Inflowing radiation	2.5	W	Inflowing radiation	1.5	W
Albedo	0	%	Albedo	30	%
End values (after 3 years):			End values (after 3 years):		
Temperature	33	°C	Temperature	-25	°C
Outgoing radiation	2.5	W	Outgoing radiation	1.1	W



From these data, André concludes that, the higher the inflowing radiation, the higher the temperature on earth. Is it correct for André to draw this conclusion? Explain your answer.

Figure 3-4 Continued

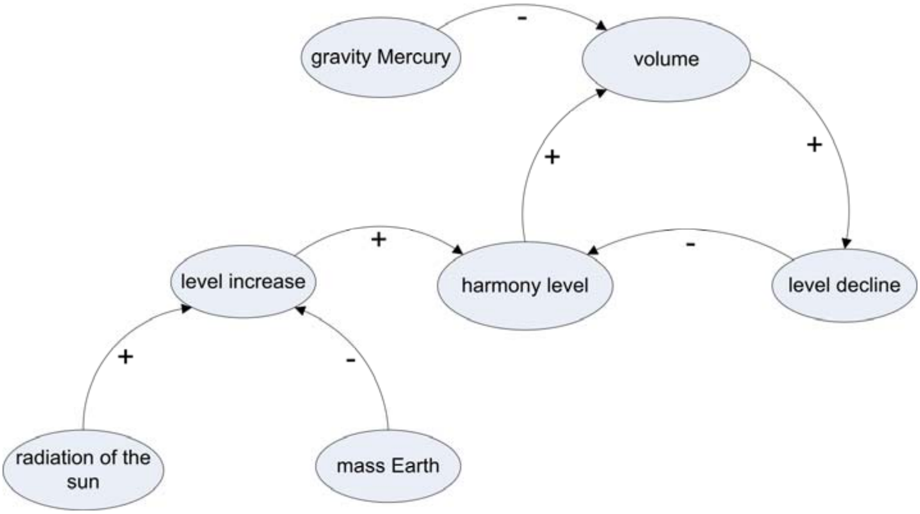


Figure 3-5 The introduction model of the harmony of the spheres that was given in the test

Apply - simple
Choose the right statement.

A. The higher Mercury's gravity, the smaller the volume.
B. The higher Mercury's gravity, the larger the volume.
C. Mercury's gravity does not influence the volume.

Explain your answer.

Apply - complex
We performed an experiment with a computer simulation. The values in the table show that the volume increased to 30 °C after three years.

Starting volume	15	dB
Mercury's gravity	375	Gals
Inflowing radiation	2.5	W
Earth's mass	75000	Et
End values (after 3 years):		
Volume	25	dB
Level decrease	2	hW

In a graph, sketch the development of the volume between start time and end time of the computer simulation. Explain your sketch. Also, describe what happens with the volume after a longer period of time.

Figure 3-6 Six harmony test items

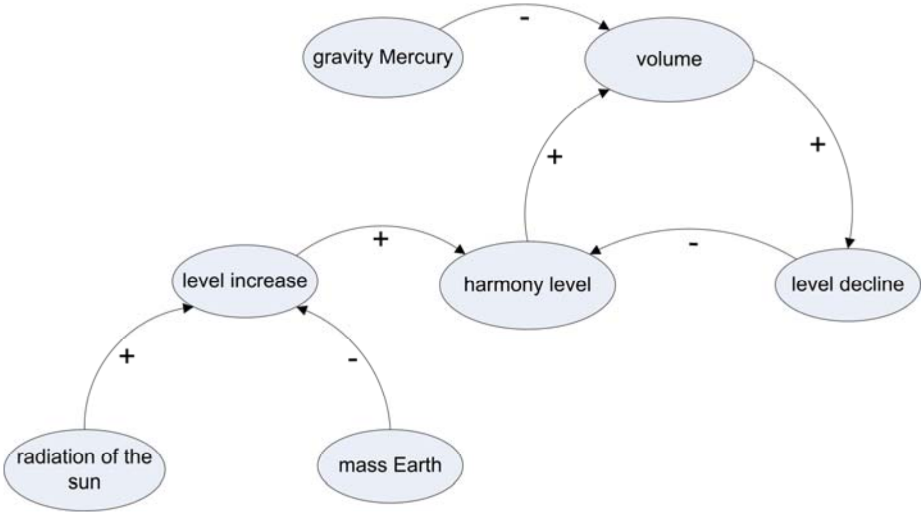
Create - simple

```
Pluto's orbit around the sun is very long and oval. Therefore, the
distance between Pluto and the sun is not always the same, but varies
from 4 to 7 billion kilometers. If Pluto is located far from the sun,
the volume turns out to be low.
```

```
Describe which new variable(s) you need to model this situation.
```

Create - complex

```
Add the phenomenon of Pluto to the following model.
```



Evaluate - simple

```
Is it correct that level increase influences the Earth's mass? Explain
why or why not.
```

Evaluate - complex

```
Is it correct that the harmony of the spheres will become deafeningly
loud after some time for a certain large amount of radiation of the
sun? Explain your answer.
```

Figure 3-6 Continued

3.4 Purpose of the Study

The purpose of this study is to validate the modeling test, in order to ensure its usefulness for classroom research. A first requirement is a reliable scoring method that enables multiple raters to agree on the judgment. For a test to be valid, it is necessary that test discriminates between the different reasoning abilities identified in the framework and between participants with different backgrounds with respect to the different dimensions.

In this study, we first examined the interrater reliability of the scoring method. Second, we analyzed construct validity of the test by looking at the structure of item responses in order to find evidence of the subskills identified in the framework. First, we expected distinguishable scales for the reasoning processes *apply*, *create* and *evaluate* and, in the case of realistic domain, reproduce conceptual knowledge. Second, the dimension of complexity may distinguish between simple and complex reasoning. In our analysis we focused on the overall effect of complexity on reasoning performance. Although this distinction may recur in the subscales for each of the reasoning processes, our analysis was limited to the overall effect, because the number of items for each reasoning process did not permit a more detailed analysis within each type of reasoning . Likewise, the effect of domain-specificity was assessed at the overall level. However, in this case, the effect of conceptual knowledge was taken into account because we expect an influence of domain knowledge on the domain-specific items.

Finally, we investigated the power of the test to discriminate between students with and without modeling experience, and between students with and without prior domain knowledge. We expected both experienced and non-experienced students and students with and without prior domain knowledge to provide a large variety in the answers. Our hypothesis is that the experienced students will perform better than the non-experienced students on all aspects of the test. Overall, we expected the group with relevant prior domain knowledge to perform better than the group without prior knowledge on the domain-specific part of the test. Furthermore, we expected students with more developed academic reasoning skills to perform better than students with limited academic reasoning skills on the domain-general part of the test.

3.5 Methods

3.5.1 Participants

The test was administered to 131 participants of three different backgrounds: 43 eleventh grade students in secondary school with a science major (12 female, 31 male; between 15 and 19 years old), 31 first-year university students of psychology (18 female, 13 male; between 18 and 25 years old), and 57 first-year university students of engineering physics who had completed the course Dynamic Modeling and Simulation prior to participating in our study (7 female, 50 male; between 18 and 21 years old). The participants were rewarded for their participation. The eleventh grade students were awarded a gift voucher worth €7,50; the first-year psychology students received course credits; and the physics students could earn bonus points for their regular exam.

3.5.2 Test

3.5.2.1 Scoring Method

For the scoring of items we identified typically correct and incorrect answer elements. In order to achieve a systematic scoring key, we searched for recurring patterns in the answer elements. A central element that occurs in all types of test items is the relation. A relation can be described at different levels: first, it can be mentioned that a relation between two particular variables exists; second, the causal direction of the relation can be specified; and finally, the nature of the relation (positive or negative) can be specified. For example, the textual expression 'the greater the radiation of the sun, the greater the amount of energy flowing to the earth' contains all three aspects of the relation between the variables ‘radiation of the sun’ and ‘inflow of energy to the earth’. The expression 'energy in the earth is needed for the temperature' is less specific and only contains the first two aspects, namely the existence and the direction of the relation, but not the third aspect of quality. See the Appendix for on overview of the standardized correct answer elements in the domain of the energy of the earth.

For each item in both subtests a participant received 0 to m points, where m + 1 was the number of answer categories. The number of answer categories varied from 2 for the least elaborate item to 5 for the most elaborate items. The incorrect answer elements reflect several types of errors: general, definitional, relational, evaluative, and creational. Figure 3-7 shows an example of the scoring method for a simple create item.

3.5.3 Procedure

The data was collected at three secondary schools and a university. All participants were given two hours to complete the test.

Before starting the test, participants were informed about the study. The eleventh grade students and the psychology students were asked not to worry about the items that were too difficult, to answer to their best ability, to keep track of time, to try to finish all items, and to work in silence. The first-year physics students were informed that the test was meant for novice modelers and that the test used the notation of a causal concept map that was more informal than they were used to in the course of dynamic modeling and simulation. The physics students completed the test immediately after the regular exam for their course.

Create - simple

Suppose, the watch has besides solar cells a *kinetic mechanism* that can produce energy. Movements of the arm and wrist will cause the kinetic mechanism to charge the battery.

Describe which new variable(s) you need to model this situation.

Answer elements:
correct
   1) kinetic energy
   2) number of movements

incorrect
   1) non-specific variable name (good intention but unclear, for example 'kinetic mechanism')
   2) non-relevant or redundant variable

Credits:
full credit
   correct element 1 or correct element 2
partial credit
   incorrect element 1
no credit
   other responses

Figure 3-7 Example of the scoring method for a simple create item with 3 answer categories

The modeling test started with a short introduction to the topic of global warming and a short introduction to the notation of causal concept maps. After completing the black sphere subtest, participants continued with the harmony subtest. The black sphere subtest was taken away immediately after completion, to avoid looking back at similar items in the other domain.

3.5.3.1 Data Analysis

Interrater reliability was analyzed by calculating Krippendorff's alpha, a reliability measure that is robust with respect to the number of observers, sample size, and presence or absence of missing data (Hayes & Krippendorff, 2007). Krippendorff's alpha is interpreted in the same way as Cronbach's alpha.

The test was intended to assess students' proficiency in four different thinking processes: the three types of reasoning apply, create, and evaluate, and reproducing conceptual knowledge. A first requirement is that the items within a single scale assess the same underlying ability. In order to investigate this aspect of construct validity, we used item response models for each of the reasoning processes (Embretson & Reise, 2000). The data consisted of 52 item responses of 131 students. Because the test items are polytomous with maximum scores ranging from 2 to 5, the item scores cannot be assumed to follow a normal distribution. These qualifications make the data less suitable for a factor analysis for which the data must be (near) normally distributed. Item response models are more flexible with respect to the distribution of item values, it can handle categorical data, and it is less restrictive to the number of items.

Item response theory provides a way to test the assumption that a set of items measures a latent ability (a so-called latent trait). Item response theory makes the assumption that the differences in item responses among the participants are explained by a different level of the latent ability. The theory uses probabilistic models that describe the mathematical relationship between an ability (latent trait) and item responses. An item response model describes one or more latent traits and gives information about the goodness of fit of the individual items that are supposed to measure the latent trait, for example the reasoning ability of evaluating.

Several parameters are estimated in an item response model, for example difficulty, discrimination or guessing parameters. If a set of test items assesses a single latent trait indeed, the model will provide a good fit with actual student scores. If the set of items draws on multiple underlying abilities, the fit will be poor and, given sufficient statistical power, the model can be rejected. Once the item parameters have been determined from the model, the model can be used to compute an ability score (theta) for each individual student. The advantage of using this ability score over the more usual sum of item scores is that the theta scores potentially are more differentiating. The estimation of theta may not have a linear relationship with the sum score.

In the one-parameter Rasch model only one parameter will be estimated for each item, namely item difficulty. The chance of a particular answer is related to the ability of a student and to the item difficulty described by the difficulty parameter. The Rasch model is a relatively simple item response model and in its simplicity it has proven to be sufficient for producing measures (Liu & Boone, 2006). Therefore, to find evidence of the structure of our modeling knowledge framework and to scale the items for the three types of reasoning and for reproducing conceptual knowledge items, we made use of the one parameter logistic model (OPLM), a polytomous extension of the Rasch model (Verhelst, Glas, & Verstralen, 1995), for each of the scales.

To investigate the discriminative power of the test with respect to participants of different levels of ability, we analyzed the variance of the scores of the three groups of participants. We used estimated thetas in the analysis of variance of the scores in the different segments of the framework and sum scores for the other (sub)parts of the framework for which no theta estimates were available. The sum scores were calculated by awarding each item with a maximum of 1 point. This means each item is given the same weight and thereby the same importance in the sum score. Partial credit was given for partly correct answers according to the item's answer categories. Therefore, in the sum score analysis the maximum score on the black sphere subtest was 29 and on the harmony subtest 23.

3.6 Results

3.6.1 Interrater Reliability

Part of the test data, 21 out of 131 tests equally distributed among the groups of students and randomly chosen, was scored by the first author and a second independent rater who was trained in using the coding scheme. Interrater agreement was examined at the level of answer elements and resulted in an acceptable interrater reliability of Krippendorff’s alpha .77. The remaining test data was coded by the first author alone.

3.6.2 Construct Validity of the Scales

3.6.2.1 Item response models

We analyzed the items for the subscales of the three types of reasoning and reproducing conceptual knowledge with one-parameter item response models (Verhelst & Glas, 1995) as implemented in the software package OPLM (Verhelst et al., 1995). We first checked whether all items together could be said to assess a single underlying ability. Although the reliability of the combined scale is high (alpha = .85), this model could clearly be rejected (R1c = 457.810, df = 384, p = .0052). Thus, although the combined scales could serve as a reliable instrument to assess overall modeling ability, the OPLM-analysis indicates that all items together do not represent a unidimensional construct and that there are multiple underlying abilities. Therefore, as a next step, we used four models, one for reproducing conceptual knowledge, and one for each type of reasoning process (still taking domain-specific and domain-general items together). This way, the resulting models could not be rejected (see Table 3-2), thus indicating that the results can be explained in terms of four one-dimensional latent abilities. Reliabilities are Cronbach’s alpha of .64 (Apply), .78 (Create), .45 (Evaluate), and .36 (Reproducing conceptual knowledge) (see Table 3-2).

Table 3-2 Statistics of the one-parameter item response models for the three types of reasoning and reproducing conceptual knowledge

Type of reasoning	mean	(SD)	max	number of items	alpha	R1c	df	p
Reproduce	7.79	(2.64)	13	6	.36	20.24	24	.68
Apply	14.71	(5.73)	37	14	.64	119.99	108	.20
Create	27.44	(8.00)	48	18	.78	138.65	141	.54
Evaluate	13.78	(4.06)	31	14	.45	81.48	90	.73

To investigate the construct validity for the framework dimension of complexity, we analyzed the difficulty of the test items for the two levels of simple and complex. The item difficulty was estimated by ConstructMap software (Kennedy, Wilson, Draney, Tutunciyan, & Vorp, 2008) that estimates item difficulty on the scale of estimated ability. The simple items had a smaller mean item difficulty of -.371 (SD = .898). The complex items had a larger mean item difficulty of .251 (SD = .840). Analysis of variance (ANOVA) of the mean item difficulty for the simple and complex items showed significant differences, $F(2, 50) = 6.493$, $p = .014$, effect size $d = .72$.

3.6.3 Comparing Group Scores

3.6.3.1 *The effect of the variables gender and age*

To ensure that the variables gender and age did not have effect on the test performance we performed the following tests. To test for gender effects, we ran a one-way ANOVA on the overall scores, $F(1, 129) = 1.53$, $p = .22$. Because this test revealed no gender differences, the gender variable could be omitted from the further analyses. To test for the effect of age, we examined the correlation between the overall test score and age. The Pearson correlation was - .15 and was nonsignificant.

3.6.3.2 *Differences for the Types of Reasoning using Estimated Thetas*

The OPLM item response models provided estimated thetas for each participant for Reproducing conceptual knowledge and for each of the types of reasoning (Apply, Create, and Evaluate). We performed analyses of variance with the estimated thetas to check the discriminative power of these scales with respect to participants with different levels of modeling proficiency. Significant differences between groups were found for all four scales (see Table 3-3).

For the Reproduce items ($F(2, 128) = 20.145$, $p = .000$), intergroup comparisons using the Scheffé post hoc criterion for significance indicated that the physics students scored significantly higher than both the eleventh grade and psychology students ($p = .001$, effect size $d = .83$; $p = .000$, effect size $d = 1.31$

resp.). There was a marginally significant difference between the eleventh grade students and the psychology students, where the eleventh grade students had the higher scores (p = .051, effect size d = 0.50).

For the Apply (F(2, 128) = 9.013, p = .000), intergroup comparisons using the Scheffé post hoc criterion for significance indicated that the physics students scored significantly higher than both the eleventh grade and psychology students (p = .035, effect size d = .53; p = .000, effect size d = .95 resp.). No difference was found between the eleventh grade students and the psychology students.

For the Create items (F(2, 128) = 6.526, p = .002), intergroup comparisons using the Scheffé post hoc criterion for significance indicated that the physics students scored significantly higher than both the eleventh grade and psychology students (p = .015, effect size d = .61; p = .010, effect size d = .68 resp.). No difference was found between the eleventh grade students and the psychology students.

For the Evaluate items (F(2, 128) = 3.751, p = .026), intergroup comparisons using the Scheffé post hoc criterion for significance indicated that the physics students scored significantly higher at $p < 0.1$ than both the eleventh grade and psychology students (p = .076, effect size d = .48; p = .080, effect size d = .48 resp.). No difference was found between the eleventh grade students and the psychology students.

Table 3-3 Means and standard deviations of the estimated thetas for the three groups of students

	Condition					
	Secondary education (n = 43)		Psychology students (n = 31)		Physics students (n = 57)	
	Mean θ	(SD)	Mean θ	(SD)	Mean θ	(SD)
reproduce score	0.146_{a}*	(0.510)	-0.148_{b}*	(0.597)	0.546_{c}	(0.450)
apply score	-0.344_{a}	(0.518)	-0.523_{a}	(0.491)	-0.097_{b}	(0.407)
create score	-0.020_{a}	(0.575)	-0.071_{a}	(0.614)	0.313_{b}	(0.514)
evaluate score	-0.595_{a}*	(0.390)	-0.614_{a}*	(0.476)	-0.388_{b}*	(0.469)

Note. Means in the same row with different subscripts differ at $p < .05$ in the analysis of variance.
*$p < 0.1$.

3.6.3.3 Differences between the domain-specific and domain-general subtests

To investigate differences between the domain-specific and domain-general subtests, we analyzed the performance of the three groups by comparing the sum (sub)scores for the two subtests separately (see Table 3-4). Because the thetas estimated the ability based on items in both subtests, we used sum scores in this analysis. The results of the group comparisons are in line with the above described results of the theta analyses. Additionally, some trends were found in the comparison of eleventh grade students and psychology students; the eleventh grade student performed better on the items for reproducing conceptual knowledge.

We analyzed the total sum scores for the subtest in the black sphere domain. A test of homogeneity of variances indicated that the variances of the three groups were not equal. Therefore, we performed a Mann-Whitney test that showed that physics students scored significantly higher than both eleventh grade and psychology students ($U = 756$, $p = .001$, effect size $d = .76$; $U = 304.500$, $p = .000$, effect size $d = 1.33$ resp.). In line with expectations, a trend was found in favor of the eleventh grade students compared to the psychology students.

An analysis of variance of the mean sum score for the black sphere apply items of the three groups of students showed significant differences ($F(2, 128) = 12.68$, $p = .000$). Intergroup comparisons using the Scheffé post hoc criterion for significance indicated that the physics students scored significantly higher than both the eleventh grade and psychology students ($p = .029$, effect size $d = .54$; $p = .000$, effect size $d = 1.15$ resp.). Again, a trend was found in favor of the eleventh grade students compared to the psychology students. The eleventh grade student (95% confidence interval [2.94, 3.69]) performed better than the psychology students (95% confidence interval [2.25, 3.06]).

Table 3-4 Means and standard deviations of the sum scores on subparts of the domain-specific black sphere and domain-general harmony test for the three groups of students

	Condition					
	Secondary education (*n* = 43)		Psychology students (*n* = 31)		Physics students (*n* = 57)	
	Mean	(SD)	Mean	(SD)	Mean	(SD)
total score	25.06$_a$	(6.50)	22.74$_a$	(6.71)	29.36$_b$	(5.03)
black sphere score	14.49$_a$	(3.93)	12.37$_a$	(4.10)	17.11$_b$	(2.92)
harmony score	10.57$_a$	(3.41)	10.37$_a$	(3.43)	12.24$_b$	(3.20)
black sphere						
reproduce score	3.39$_a$	(1.19)	2.67$_b$	(1.21)	4.25$_c$	(0.93)
apply score	3.31$_a$	(1.22)	2.66$_a$	(1.10)	3.95$_b$	(1.15)
create score	4.99	(1.99)	4.49$_a$	(2.16)	5.77$_b$	(1.61)
evaluate score	2.80	(0.93)	2.55$_a$	(1.21)	3.15$_b$	(1.05)
harmony						
apply score	2.50	(1.43)	2.19$_a$	(1.32)	3.01$_b$	(1.36)
create score	5.00$_a$	(1.60)	5.09	(1.70)	5.88$_b$	(1.42)
evaluate score	3.07	(1.31)	3.20	(1.38)	3.65	(1.16)

Note. Means in the same row with different subscripts differ at $p < .05$ in the analysis of variance.

3.7 Conclusion and Discussion

In this study we investigated the validity of a modeling knowledge test based on the ACE framework of modeling knowledge. Using item response models, we found evidence of four distinguishable abilities, corresponding with the reasoning processes proposed in the ACE-framework. Reliabilities of the Create and Apply scales were acceptable, whereas the reliabilities of the Reproduce and Evaluate scales were low. Although, the low reliabilities on parts of the test limit its power to detect small differences, the differences between the groups in our study could still be detected.

The results of the analysis of variance show differences between the students with and without modeling experience as expected. The physics students performed significantly better than the other two groups on almost all parts of the test. Moreover, domain knowledge seemed to play a role in the test performance. In the black sphere subtest, comparing the eleventh grade and psychology students we found a trend in favor of the eleventh grade students. The difference between these groups on the Reproduce items was significant. In the harmony subtest the scores of eleventh grade students and psychology students show no differences. The high performance of eleventh grade students on the black sphere subtest can be explained by the actual presence of domain knowledge whereas the psychology students did not have domain knowledge. We expected the psychology students to perform better than the eleventh grade students on the domain-general items based on a better development of general reasoning skills. However, the results show no significant differences on the harmony test. This may suggest that model-based reasoning abilities are not simply developed hand in hand with general academic reasoning abilities. Another explanation may be that the test is not sensitive enough to detect differences between eleventh grade students and first-year psychology students. Further study should clarify this.

Our study has some limitations. First, two of the four scales appeared to have a low reliability. This makes these scales less suitable to be used as single tests standing on their own and to detect individual differences. Second, there is a tension between the research goal of validating the test and educational practice. To validate each single cell in the framework, we need to develop more items in each cell. In this study, we first focused on the main dimension of the types of reasoning and on the dimension of complexity. The dimension of domain-specificity still needs to be further investigated and will especially become relevant when after a larger amount of modeling experience the model-based reasoning skills are generalized to other domains.

For assessment purposes in educational practice, the research instrument can be made more efficient. If there is no need to assess all subscales independently, the test length could be reduced considerably, while maintaining a balanced distribution of test items over the different aspects of model-based reasoning. Furthermore, if the test is only used as a posttest, it could be adapted to use the model representations that have been taught in class.

In conclusion, the framework proved to be a useful tool to operationalize the specific aspects of modeling knowledge in a test in the domain of global warming. This study provided evidence that apply, create, and evaluate are distinguishable processes in model-based learning outcomes. The ACE test operationalized the processes and made it possible to develop instruction according to the specific reasoning processes. Future research may use the test in an experimental setting using the tests in the two different domains as pre and posttests. Another important result is that the framework enables the development of test items in other domains. Generic guidelines followed from the test item development in the domain we described in this study and these guidelines can be applied in other domains provided the domain structure contains a sufficient level of complexity. Furthermore, the framework may also be used for designing an observation scheme to interpret students' modeling performance. As such we think the framework will be an adequate tool to evaluate the learning of computer modeling skills and help the field in studying the development of knowledge under varying conditions of modeling.

Chapter 4 What can be learned from computer modeling? Comparing expository and modeling approaches to learning dynamic systems behavior

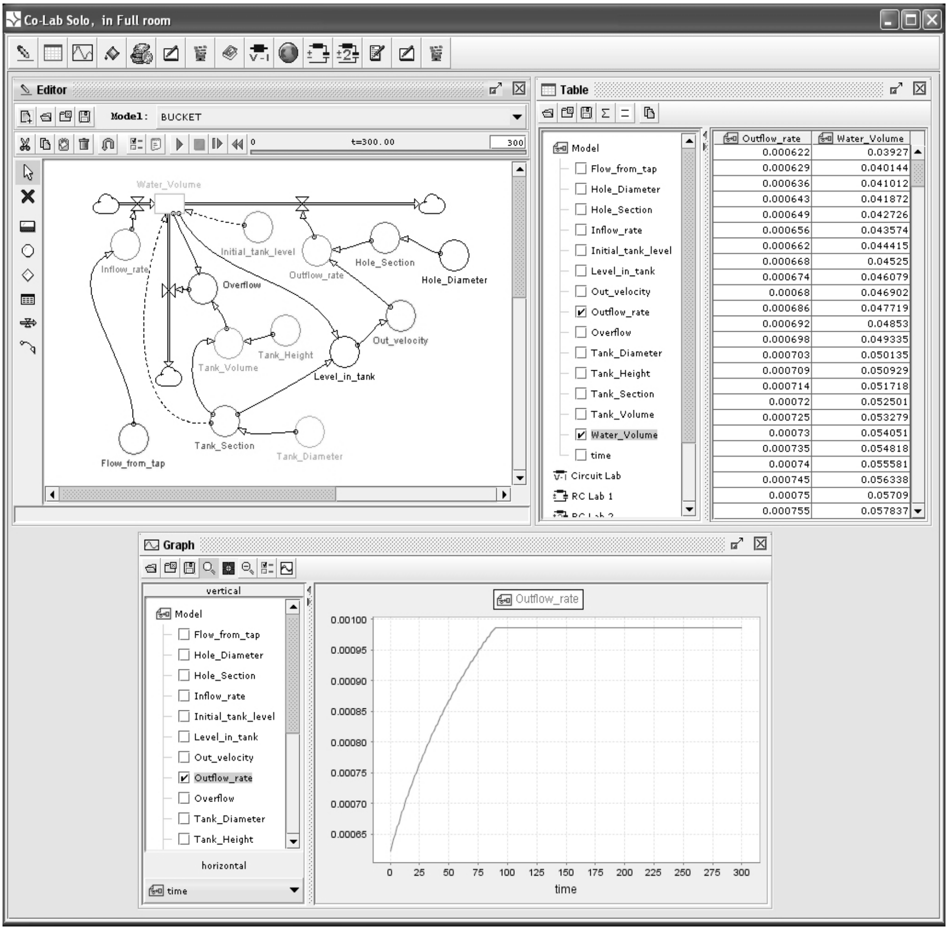
Abstract

In this chapter, we compare the learning outcomes of two instructional approaches, expository teaching and computer modeling. For this we use the ACE test, an assessment instrument specifically aimed at the knowledge and skills that should be acquired using computer modeling. A group of students who performed a computer modeling task in the domain of global warming was compared with a group who was taught traditionally. The assessment aimed to discriminate between different types of learning outcomes based on the ACE modeling knowledge framework. We found differences in learning outcomes between the two instructional approaches on several scales of the test. While both groups performed equally well on simple problems, the modeling group outperformed the expository group on complex problems. More specifically, the modeling group outperformed the expository group on problems that required them to reproduce complex conceptual knowledge and evaluate complex models and data. No differences were found for items that required the application of knowledge or the creation of models. These results show that the ACE test is sensitive in revealing differences for specific parts of the ACE framework between the two instructional approaches and has discriminative power with respect to the two instructional approaches.

4.1 Introduction

Computer modeling is an emerging topic in secondary science education, in which students construct or modify models of dynamic systems that can be simulated. Through the construction of models and experimentation with the resulting simulations, learners are supported in building understanding about complex dynamic systems. Despite the fact that modeling of dynamic systems appears to be very difficult for secondary education students (Cronin & Gonzalez, 2007; Fretz et al., 2002; Hmelo, Holton, & Kolodner, 2000; Sins et al., 2005; Sterman, 2002; Wilensky & Resnick, 1999), its expected benefits make it a worthwhile activity in the science curriculum (Magnani et al., 1998; Mandinach, 1989; Stratford et al., 1998). Modeling dynamic systems invokes specific learning outcomes compared to other modes of learning and these learning outcomes need to be specified in order to make them measurable.

Computer modeling finds a place in approaches that see learners as active constructors of knowledge, in which case the construction of knowledge is materialized as the construction of an interactive external representation. Learners investigate and experiment with a phenomenon from the real world in order to create a model. Also, they investigate and experiment with the model they constructed (Doerr, 1997; Hennessy et al., 2007). Models constructed by learners can play a role in the process of inquiry as explicit representations of learners' ideas. For instance, parts of a model under construction may represent a learner's hypothesis about (a fragment of) the observed phenomenon (Van Joolingen & De Jong, 1997). An example of a computer modeling environment is shown in Figure 4-1. This modeling environment Co-Lab provides a modeling, table and graph tool and will be used in this study.



Outflow_rate	Water_Volume
0.000622	0.03927
0.000629	0.040144
0.000636	0.041012
0.000643	0.041872
0.000649	0.042726
0.000656	0.043574
0.000662	0.044415
0.000668	0.04525
0.000674	0.046079
0.00068	0.046902
0.000686	0.047719
0.000692	0.04853
0.000698	0.049335
0.000703	0.050135
0.000709	0.050929
0.000714	0.051718
0.00072	0.052501
0.000725	0.053279
0.00073	0.054051
0.000735	0.054818
0.00074	0.055581
0.000745	0.056338
0.00075	0.05709
0.000755	0.057837

Figure 4-1 The learning environment Co-Lab with its modeling tool (top left window)

Recently, the debate about the effectiveness of constructivist and in particular inquiry approaches to learning has gained attention (Kirschner et al., 2006; Klahr & Nigam, 2004; Rittle-Johnson & Star, 2007). The tendency of the criticism regarding inquiry learning is that in experimental studies "unguided instruction" has shown no benefits when compared with "direct instruction". For example, in a larger scale international study Lederman, Lederman, Wickman, and Lager-Nyqvist (2007) compared inquiry learning with direct instruction and found no difference in learning gains for the two contrasting methods of instruction. The best instructional method might be a mix of inquiry learning and direct instruction (see also Mayer, 2004). In a study by Star and Rittle-Johnson (2008), discovery learning in which prompts were given about possible strategies and direct instruction about the different strategies were compared in the learning of solving linear equations. On the one hand, it appeared that both modes of instruction led to learning gains, i.e., an improvement of flexibility in problem solving, and that both instructional approaches were compatible. On the other hand, studies show the benefits of inquiry learning on inquiry-specific learning outcomes such as process skills (Geier et al., 2008), 'more sophisticated reasoning abilities' involved in solving complex, realistic problems (Hickey, Kindfield, Horwitz, & Christie, 1999), and scientific thinking skills in guided inquiry (Lynch, Kuipers, Pyke, & Szesze, 2005). Benefits of inquiry learning have also been found with respect to learning goals that are similarly covered by direct instruction such as understanding of science content (Geier et al., 2008).

The claim that inquiry instruction has no added-value compared to direct instruction can be challenged at various points. Learning gains are involved in both inquiry and expository approaches of teaching, but the specific differences and similarities are hard to measure and require a specific test instrument. Moreover, the modes of instruction can be implemented in various ways: inquiry learning can be guided by more or less support and direct instruction can be implemented with more or less opportunity for discovery.

In the current chapter, we address the important issue when comparing different modes of instruction: how to measure the learning outcomes of computer modeling. In comparing two approaches to learning one should devise a test for learning that does right to the claims of these approaches. For instance, when evaluating an approach that claims to improve the ability to solve calculus problems, one should evaluate learners' calculus problem solving ability. Therefore, in order to evaluate computer modeling, it is important to investigate its claims with respect to learning outcomes and then to create instruments to test these outcomes.

Various benefits of computer modeling have been claimed in the literature. First, modeling is a method for understanding the behavior and characteristics of complex dynamic systems (Booth Sweeney & Sterman, 2007; Sterman, 1994). Second, modeling is assumed to enhance the acquisition of conceptual knowledge of the domain involved (Clement, 2000). Modeling has the potential to help learners develop high-level cognitive processes and thereby to facilitate conceptual change (Doerr, 1997). Third, modeling is assumed to be especially helpful for the learning of scientific reasoning skills (Buckley et al., 2004; Mandinach & Cline, 1996). Key model-based scientific reasoning processes are described by *creating*, *evaluating*, and *applying* models in concrete situations (Wells et al., 1995).

In order to assess the model-based learning outcomes we developed the ACE (Apply Create Evaluate) test based on the ACE framework, that provides a precise operationalization of different types of modeling knowledge and skills (Van Borkulo et al., 2008). The ACE framework served as the basis for the development of the model-based reasoning test that was validated in a previous study .

In the current chapter, we present a comparative study with two contrasting modes of instruction: expository instruction and computer modeling. As described previously, several studies argue that there is no added-value of inquiry instruction over direct instruction. Yet, we claim that modeling-based learning involves heterogeneous activities that are partly uniquely related to a modeling task and should be assessed accordingly, and that are partly also involved in expository teaching. Therefore, we expect specific differences on the model-based reasoning processes of applying, creating, and evaluating models when comparing modeling and expository teaching, but also similarities. In the view of this study, the expository mode of instruction directly exposes the information to the learners in textual format, providing guidance in the form of assignments, but without any dynamic tools such as simulations or concept maps and without explicit model building. The modeling mode of instruction comprises a guided inquiry approach supported by modeling and simulation tools. The two modes of instruction were compared using the ACE test, a test that is dedicated to detect specific knowledge gained by modeling activities. Before describing the details of the study, we will briefly summarize the ACE framework.

4.2 Modeling knowledge: ACE framework

The ACE framework, as described in van Borkulo et al. (Van Borkulo et al., 2008), distinguishes three dimensions of knowledge: 1) types of reasoning, 2) complexity, and 3) domain-specificity (see Figure 4-2). The first dimension comprises the core reasoning activities of a modeling activity: *applying* (A) knowledge of relations in a model by making predictions and giving explanations, *creating* (C) variables and relations between variables into a model, and *evaluating* (E) models and experimental data produced by a model.

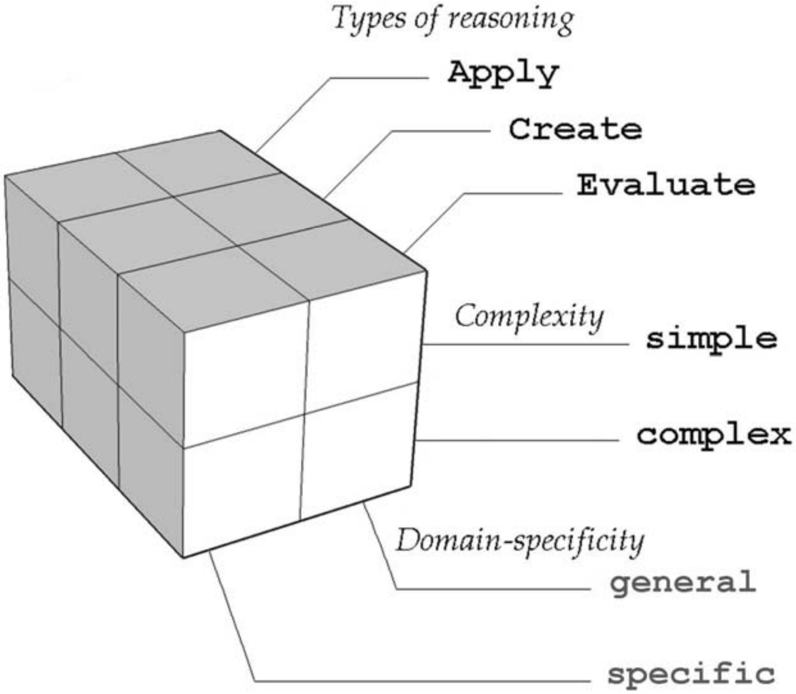
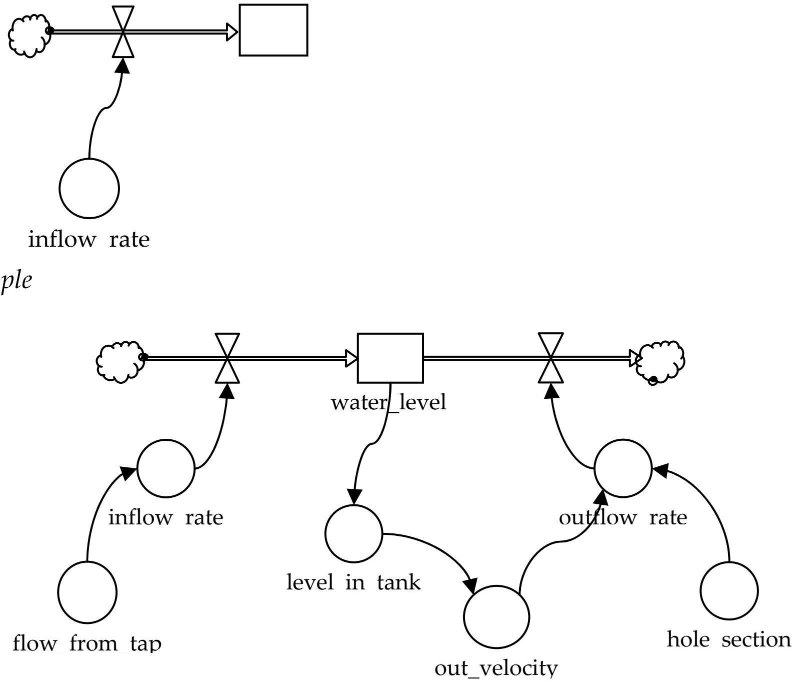


Figure 4-2 Modeling knowledge ACE framework

The second dimension concerns the aspect of complexity. Modeling is typically used to understand complex dynamic systems and understanding complex systems is fundamental to understand science (Assaraf & Orion, 2005; Hagmayer & Waldmann, 2000; Hmelo-Silver, Marathe, & Liu, 2007; Hogan & Thomas, 2001; Jacobson & Wilensky, 2006). In the framework we distinguish *simple* and *complex* model units based on the number of variables and relations involved. A simple unit is the smallest meaningful unit of a model, with one dependent variable and direct relations to that variable only, and a *complex* unit is a larger chunk that contains indirect relations and possibly (multiple) loops and complex behavior (see Figure 4-3). Since the derivation of indirect relations in a causal network is often complex and computationally more demanding (Glymour & Cooper, 1999), a test item about indirect relations will invoke more complex reasoning.



Simple

Complex

Figure 4-3 Example of a simple and a complex model part

The third dimension covers the aspect of domain-specificity. Modeling invokes the acquisition of domain knowledge. By creating a model and experimenting with the model, learners gain insight in the concepts and structures within the domain. Furthermore, reasoning with the created model is influenced by the availability of relevant prior domain knowledge (Fiddick et al., 2000). Therefore, reasoning may be different in a familiar than in an unfamiliar domain. In an unfamiliar domain, the only information learners have is the model itself. Reasoning must take place following the relations in the model in a step-by step way, building a reasoning chain. In a familiar domain, learners may bypass part of these reasoning chains because they remember the outcome of the chain as a whole. For instance, in a model that includes a capacitor, a person with knowledge of electronics will be able to reason that the voltage over the capacitor will increase as a consequence of a charging current, stepping over the charge as an intermediate variable. In an unfamiliar domain such a reasoning shortcut will not be possible.

4.3 Research question

In this chapter, the two contrasting instructional approaches of modeling and expository teaching are compared and the differences are investigated in the specific model-based learning outcomes between the groups related to the dimensions in the ACE framework.

We expected differences in learning outcomes on several aspects. Because the modelers had tools that supported the creation and exploration of conceptual structures with a concrete artifact, we expected the modelers to display better reasoning with the complex models. The runnability of their own models and the availability of a simulation tool enabled the modelers to perform experiments and to evaluate experimental data. Moreover, a large part of evaluating experiments is making predictions and thereby applying the rules of system dynamics by reasoning with the relations. Therefore, we expected the modelers to perform better on the subscales that measure the reasoning processes *evaluate* and *apply*. Furthermore, a substantial amount of time will be spent on constructive activities such as translating concepts into functional variables and creating relations between variables. Thus, we expected differences in favor of the modelers on the *create* scale. Finally, the expository learners are more directly and explicitly exposed to the concepts in the domain. Therefore, we expected the expository learners to be more efficient in *reproducing* conceptual domain knowledge.

4.4 Method

4.4.1 Participants

Participants were eleventh grade students from two schools. The participants were between 16 and 19 years old (M = 17.20, SD = .55) and all followed a science major. In total 74 participants completed the instruction; 68.9 % were male, 31.1 % were female.

4.4.2 Materials

4.4.2.1 Co-Lab learning environment

The Co-Lab software (van Joolingen et al., 2005) provided a learning environment for each of the two conditions to learn about global warming. One environment was designed for modeling-based instruction and consisted of a simulation of the basic energy model of the Earth, a modeling editor to create and simulate models, graphs and tables to evaluate the data produced by the model, and textual information about the domain. A different environment was designed for expository instruction and consisted only of the textual information.

All participants used worksheets with assignments about the factors in global warming. The work was subdivided in three parts. The first part was about climate models in general and included questions about the quality and accuracy of using models to make global warming predictions. The second part discussed the factors albedo and heat capacity, provided information, and included questions about the influence of these factors on the temperature on Earth. This was implemented in different ways for the modelers (who created a model to support their reasoning) and the learners in the expository mode of instruction (who could only use the information provided). For example, an assignment about the influence of the albedo on the equilibrium temperature asked both groups to predict what would happen with the equilibrium temperature if the albedo was high or low respectively. Subsequently, the modelers were asked to investigate their hypotheses with their model. The third part was about evaluating one's understanding of the domain structure. The modelers were asked to compare their own model's behavior with the model of the given simulation of the phenomenon. The learners in the expository mode of instruction were asked to compare their findings about the influencing factors with given global warming scenarios. The scenarios specified a number of plausible future climates under the assumption of different values of future emissions of greenhouse gasses. As a final product the expository learners wrote a report about the factors influencing the temperature on Earth as opposed to the modelers who created a model.

4.4.3 The Modeling ACE Test

Based on the ACE framework, we developed the ACE test consisting of a domain-general and a domain-specific part (Van Borkulo et al., 2008). We used the domain-general part of the ACE test as a pretest and the domain-specific part as a posttest. The domain-general fantasy 'harmony' test assessed domain-general modeling skills before the intervention and the domain-specific global warming 'black sphere' test was used as a posttest to assess model-based reasoning skills in the instructed domain. For reasons of time constraints we chose to focus on the domain-specific reasoning processes. Therefore, in this study we analyzed the domain-specific part of the ACE test and used the domain-general ACE test as a covariate to control for individual differences in prior modeling skills.

The domain-general test was about the fictitious phenomenon of the ″harmony of the spheres″. Because it was about a fictitious phenomenon, students would not have any relevant domain knowledge, or experiential knowledge to rely on. The domain-specific test was about the domain of global warming, where students would have relevant domain knowledge after the intervention. The model structures for both tests were isomorphic.

In a previous study (see Chapter 3), the test was validated with students of different levels of modeling proficiency. The data analysis confirmed the existence of four subscales: the types of reasoning apply, create, and evaluate, and the reproduction of conceptual knowledge. Furthermore, the test showed discriminative power with respect to groups of students of different modeling experience and suggested a positive influence of domain knowledge on the test scores.

The outcomes of the validation study have led to some changes of the ACE test. Also time constraints on the current experimental setting limited the total number of items. Two apply and three evaluate items were textually modified, one apply and one evaluate item were removed, because the original version appeared to be ambiguous in its formulation, leading to bad performance in our analysis.

Also, some items appeared to be too difficult or too easy. For this reason, three create items and one evaluate item were removed because they were too difficult. Three create items were removed because they were too easy. One apply item and one evaluate item were modified because they were too easy. Finally, for reason of dependency between items three create items were removed and in one apply item variable names have been replaced by others.

In the current study, the harmony test consists of 19 items in the core ACE types of model-based reasoning in an unfamiliar domain (see Table 4-1). Additionally, to be able to control for students' prior knowledge about the modeling formalism six items were added about core elements in the modeling language: stocks, flows, and feedback: 3 simple and 3 complex items. Figure 4-4 shows the introduction model that was given in the pretest. Figure 4-5 shows examples of a simple apply item and a complex evaluate item.

Table 4-1 Distribution of the number of items in pre- and posttest among the framework dimensions

Number of items	Pretest		Posttest	
	Harmony		Black sphere	
	Simple	complex	simple	complex
Reproduce items*	3	3	3	3
Apply items	3	4	3	3
Create items	2	4	3	3
Evaluate items	3	3	3	3
	11	14	12	12
Total	25		24	

* The test items about the modeling formalism are classified under "reproduce harmony", although this is technically incorrect.

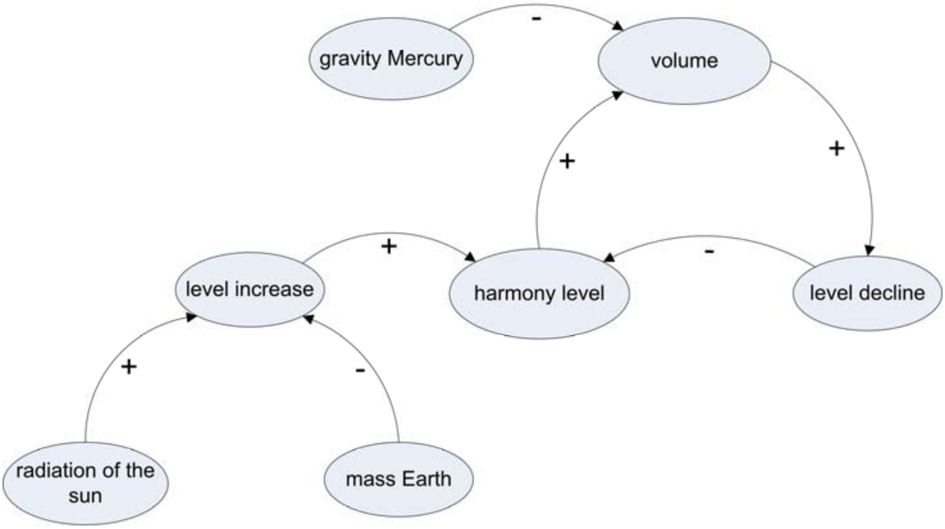


Figure 4-4 The model of the harmony of the spheres that was given in the fantasy pretest

Apply - simple

Choose the right statement.

A. The lower Mercury’s gravity, the smaller the volume.
B. The lower Mercury’s gravity, the larger the volume.
C. Mercury’s gravity does not influence the volume.

Explain your answer.

Evaluate - complex

Is it correct that a higher Mercury’s gravity will cause a faster decrease of the harmony of the spheres? Explain your answer.

Figure 4-5 Two examples of fantasy pretest items

The ‘black sphere’ test concerns modeling global warming and hence involves the domain of energy of the sun and the Earth. The black sphere test consists of in total 24 items and includes 18 items in the core ACE types of model-based reasoning (see Table 4-1). Additionally, six domain-specific items covered the reproducing of conceptual knowledge in the black sphere domain: 3 simple and 3 complex items. Figure 4-6 shows the introduction model that was given in the posttest. Figure 4-7 shows examples of a simple reproduce item and a complex create item.

In this study students' performance on parts of the black sphere test is analyzed. We are specifically interested in effects on the level of simple and complex items. Because no theta estimates are available on this level, sum scores are used in the analysis. For each item a participant received 0 to 1 points. Partial credit was given for partly correct answers. The maximum score on the harmony test was 25. The maximum score on the black sphere test was 24.

The notation used in the test was a causal concept map notation. Variables are represented by circles labeled with a variable name, causal relations are represented by arrows, and the quality of the relation is expressed by a plus or minus sign (see Figure 4-4 and Figure 4-6).

The test was offered as a paper-and-pencil test. In order to assess learners’ prior modeling ability, the harmony subtest on the fictitious domain was used as a pretest. The domain-specific black sphere subtest was used as a posttest after the instruction about global warming. The scores on the domain-generic harmony pretest were used to match participants in the experimental groups.

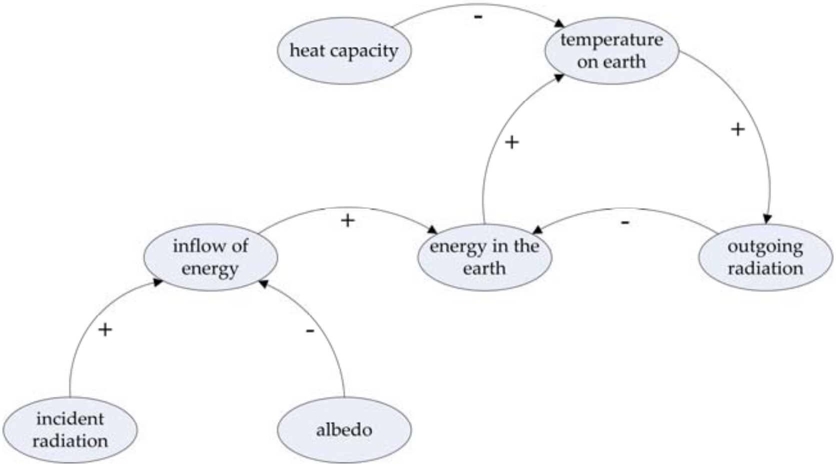


Figure 4-6 The black sphere model that was given in the global warming posttest

Reproduce - simple

What is albedo (or reflectivity) of a substance?

Create - complex

Suppose you have a watch running on solar cells. Just like a solar collector, the solar cells pick up the visible light of the sun, convert it to electricity, and store the electricity in the watch's battery. If you often use the watch's stopwatch function, the battery will empty earlier. Also, the intensity of the display's lighting is adjusted to the amount of energy in the battery. When the lighting is weak, you know the battery is almost empty.

Draw a model of this situation in order to be able to predict the empty time of the battery. Explain your drawing.

Figure 4-7 Two examples of black sphere posttest items

4.4.3.1 *Scoring method*

In the previous validation study (see Chapter 3) we developed a scoring scheme based on an analysis of the item responses of students of different levels of modeling proficiency. For each item, an answer model was derived with the elements defining the correct answer and elements representing common errors.

A frequently recurring group of elements in all subscales of the framework is the group of three elements belonging to a relation. These three elements are: an expression that indicates the *existence* of a relation, the *direction* of a relation (causality), and the *quality* of a relation (positive or negative influence). A relation can be expressed not only textually in a written explanation, but also schematically in the drawing of a model. The threefold scoring of a relation provides a detailed view of the elaborateness of students' reasoning.

4.4.4 Procedure

The experiments were held at two secondary schools. The experiment consisted of two sessions of 200 minutes each with an interval of two to four weeks depending on the school. The lessons were led by the experimenter and were additional to the regular curriculum and compulsory for each student. Participants from one school could earn extra points that were added to their regular physics grade.

All participants first attended a session in which modeling was introduced, using an example on the spreading of diseases and a leaking water bucket. After 150 minutes of introduction, the participants had 50 minutes to complete the harmony test. For the second session, the students were separated in two conditions based on equal distribution of the pretest harmony scores. For both conditions we included all combinations of school, teacher, class, and gender. In the second session, both conditions were given information and assignments about the factors influencing the temperature on Earth. In addition to the assignments, the students in the modeling condition (N = 38) performed an elaborated modeling task. The students in the expository condition (N = 36) wrote a report on the factors in global warming. After 150 minutes, all participants completed the black sphere posttest which took 50 minutes.

4.5 Results

To test our hypotheses, we computed analyses of variance with the pretest subscore as a covariate, although the scores of the participants in the two instruction modes did not differ significantly (see Table 4-2). In the analysis of black sphere *subscores,* pretest *subscores* were used as a covariate. For example, in the analysis of the scores on the simple create subtest we used the scores on the pretest simple create subtest as a covariate. For the ‘reproduce’ items this was impossible, and no covariate was used.

Table 4-2 Means and standard deviations of the pretest (sub)scores for the two conditions

		Harmony pretest				
		Expository (*n* = 36)		Modeling (*n* = 38)		max
Overall	simple	5.11	(1.39)	5.01	(1.45)	11
	complex	5.89	(2.27)	6.08	(2.46)	14
	total	11.00	(3.29)	11.09	(3.67)	25
Reproduce	simple	0.38	(0.61)	0.31	(0.52)	3
	complex	0.91	(0.60)	0.75	(0.47)	3
	total	1.29	(0.91)	1.06	(0.81)	6
Apply	simple	1.30	(0.50)	1.18	(0.65)	3
	complex	2.21	(1.22)	2.36	(1.14)	4
	total	3.51	(1.56)	3.54	(1.58)	7
Create	simple	1.68	(0.57)	1.70	(0.48)	2
	complex	1.54	(0.79)	1.51	(0.99)	4
	total	3.23	(1.18)	3.21	(1.26)	6
Evaluate	simple	1.75	(0.71)	1.82	(0.59)	3
	complex	1.24	(0.92)	1.46	(0.95)	3
	total	2.99	(1.37)	3.28	(1.24)	6

4.5.1 Harmony test and black sphere test

In line with our expectations, we found no significant main effect of condition on learning outcome, although there was a trend in favor of the modeling condition ($F(1, 72) = 2.972, p = .089$).

However, we expected differences on the subscales. When looking at the complex items overall, we found a significant difference in favor of the modeling condition ($F(1, 72) = 8.780, p = .004$, partial $\eta^2 = .110$). More specifically, students in the modeling condition performed significantly better on both the complex reproduce items ($F(1, 72) = 7.065, p = .010$, partial $\eta^2 = .089$) and the complex evaluate items ($F(1, 72) = 3.966, p = .050$, partial $\eta^2 = .053$).

For the other subscales in the framework no significant differences were found (see Table 4-3). No significant differences were found on the create subscale;

neither on the total create scores, nor on the simple create items, nor on the complex create items.

Table 4-3 Means and standard deviations of the black sphere posttest (sub)scores for the two conditions

		Black sphere posttest				
		Expository (n = 36)		Modeling (n = 38)		max
Overall	simple	6.59	(1.38)	6.58	(1.68)	12
	complex	3.67*	(1.50)	4.72*	(1.72)	12
	Total	10.26	(2.48)	11.30	(3.10)	24
Reproduce	simple	2.00	(0.80)	1.74	(0.84)	3
	complex	1.06*	(0.71)	1.50*	(0.69)	3
	total	3.06	(1.09)	3.23	(1.28)	6
Apply	simple	1.04	(0.69)	1.21	(0.70)	3
	complex	0.90	(0.69)	1.12	(0.71)	3
	total	1.95	(1.16)	2.33	(1.18)	6
Create	simple	2.09	(0.69)	2.07	(0.85)	3
	complex	1.16	(0.74)	1.26	(0.76)	3
	total	3.24	(1.34)	3.33	(1.48)	6
Evaluate	simple	1.46	(0.59)	1.56	(0.66)	3
	complex	0.54*	(0.53)	0.85*	(0.63)	3
	total	2.00	(0.79)	2.41	(0.99)	6

* Means differ at $p < .05$ in the analysis of variance.

4.6 Discussion

The aim of this study was to investigate the specific learning outcomes of computer modeling in comparison to expository instruction. The learning outcomes were examined using the domain-specific part of the ACE modeling test.

On the overall posttest score the modeling condition scored higher, but the difference was not significant. Clear differences were found with respect to the complex items. The modeling condition performed significantly better on the overall complex items. It seemed that the modelers had better strategies to cope with complex structures. More specifically, the learning gain concerned the complex evaluate items and the complex reproduce items. On the simple items in both evaluate and reproduce categories the groups performed equally well. The simple items seemed to invoke knowledge that had been acquired in equal level for both groups.

Unexpected was the difference on the complex reproduce items which was in favor of the modelers. Because the expository learners were more explicitly exposed to the domain concepts, we expected them to perform better on the reproduce items. A possible explanation is that complex conceptual knowledge to a large extent depends on structural reasoning skills and is not simply reproduced.

The simple items showed no differences between the two conditions. For the individual items in the black sphere posttest, the mean proportion correct difficulty *p* value for the simple items was .54 (SD = .22) and for the complex items .34 (SD = .12). Although there is no ceiling effect in the test scores, the results suggest that the reasoning with a model on a simple level with direct relations is equally mastered by modelers and non-modelers and that modeling training does not contribute in particular to learning gains in this field.

Against our expectations, we found no differences related to the application and creation of models. The create items in the posttest asked to model phenomena that were similar to the phenomena they practiced with. One create item presented a context that asked for transfer to a more elaborate situation. We expected the modelers to be able to perform well on the items with similar model structures. Explanations for this unexpected lack of difference include that the limited amount available for the modeling activity was too short for a difference to emerge, or that the actual behavior by students engaged in the modeling was ineffective. For instance this could be the case when instead of creating models from scratch, learners merely copied their models from given examples. For instance, during the second session a common error for the modelers was to leave out the temperature variable from the models they created. Apparently, the modelers superficially copied the familiar model structures instead of reasoning and experimenting with the model and discovering mistakes with respect to the new context. Ideally, the modelers had the opportunity to learn from their mistakes by receiving feedback from the simulation of their model as opposed to the expository learners who did not receive feedback. Thereby, the modelers had the opportunity to learn in a feedback loop: one step forward leading to another step forward. However, the feedback loop seemed to work in the reversed way: one mistake leading to another mistake.

In conclusion, the domain-specific part of the ACE test appeared to be able to detect differences in specific learning outcomes between a modeling and expository instruction concerning complex knowledge structures, with respect to reproducing complex conceptual knowledge and evaluating models. Thereby, the ACE modeling test instrument that was based on the ACE modeling knowledge framework showed discriminative power and contributes to an evidence-based discussion about the effectiveness of different modes of instruction. Moreover, the test classifies the learning outcomes of modeling in a systematic way and clarifies the learning goals of modeling and how they can be measured. However, not for all expected areas differences arose. The create test items seemed to represent achievable modeling assignments with a bounded context having a structure similar to the basic model in the instruction. Yet, the modeling condition appeared to be unable to apply the basic modeling actions they studied in the instruction in creating a model by themselves in a similar but new context. In combination with the low scores on the apply items, this suggests that the students are not used to reason every part of their model in creating and checking the functionality of the model, and that the overall modeling proficiency can be improved.

Relational reasoning seems to be an important factor in creating and evaluating a model. Applying knowledge of a model is not obviously involved in creating a relation. In this study, the participants were creating relations, but seemed not to learn how to reason. It is worthwhile to further investigate how the acquisition of create skills can be supported and how the support for the different parts of create skills can be implemented in the instruction. Future research will need to focus on an effective implementation.

Chapter 5 In search of computer modeling learning gains: comparing modeling and simulation inquiry learning

Abstract

In this chapter we compare the learning outcomes of two closely related instructional approaches: computer modeling and simulation-based instruction. The ACE test is used to compare the performance in model-based reasoning processes of a group of students who performed a computer modeling task to a group who performed assignments using a ‘ready-made’ simulation. The aim of the study was to investigate the discriminative power of the ACE test with regard to the different types of learning outcomes as described in the ACE modeling knowledge framework. We found differences in learning outcomes between the two instructional approaches on several aspects of the test. The modeling group outperformed the simulation inquiry group on complex problems and both groups performed equally well on simple problems. In particular, students in the modeling group outperformed the simulation inquiry group on problems that required them to apply reasoning in complex model structures as well as in problems requiring them to create simple models. The simulation inquiry group performed significantly better on reproducing simple conceptual knowledge. No differences were found for items that required evaluation of models and data. These results show that the ACE test is sensitive in revealing differences for specific parts of the ACE framework between two closely related instructional approaches. As expected, evaluating skills are equally developed in both approaches. The test appears to have discriminative power not only with respect to contrasting instructional approaches (see Chapter 4) but also with respect to closer related approaches.

5.1 Introduction

Despite the growing interest in modeling-based instruction, there has been relatively little analysis of the specific cognitive learning outcomes of modeling. From the studies that have investigated such outcomes (Booth Sweeney & Sterman, 2000; Kainz & Ossimitz, 2002; Komis, Ergazaki, & Zogza, 2007; Schwarz & White, 2005; Wilensky & Resnick, 1999), we have learned about the aspects of modeling. However, the various theories about the cognitive learning outcomes are developed in different contexts and the experimental results of the studies are difficult to compare. What is lacking is a comprehensive framework for classifying cognitive learning outcomes that combines multiple aspects of model-based learning and that provides enough fine-grained detail to make a distinction between closely related modes of instruction.

To fulfill the need of a framework that describes the cognitive learning outcomes of modeling, we developed the ACE framework (see Figure 5-1), focusing on the reasoning activities of applying (A), creating C), and evaluating (E). The framework contains three dimensions: first, *types of reasoning*, i.e., apply, create, and evaluate; second, *complexity*, i.e., simple and complex; and, third, *domain-specificity*, i.e., domain-dependent and domain-general. For an expanded definition of this framework, see Chapter 2.

In the previous chapter we used a test based on the ACE framework to compare modeling with expository instruction, with as a main result that modelers scored best on items measuring knowledge about complex structures. In the current chapter we zoom in on the differences between modeling-based and simulation-based inquiry learning. Computer modeling and learning with simulations are closely related (Doerr, 1997), meaning that this will provide a more rigorous test for our framework. A 'ready-made' simulation of a phenomenon can be used to experiment and to evaluate the underlying model. The similarities of modeling and simulation learning are grounded in the fact that a computer model is runnable and therefore can be used to simulate the modeled phenomenon. In fact, computer modeling is a self-made simulation, in which the variables and relations are defined by the modeler. The added value of modeling over learning from a ready-made simulation of a phenomenon is that students have to make a concrete and active representation of concepts which helps to create a mental model and to gain insight into the structure of the domain (Hestenes, 1987; Spector, 2000). Therefore, modeling can be said to be *expressive* as opposed to *explorative* for simulation based learning, following Bliss et al. (1992). The activity of creating a model is combined with the evaluation of the model by doing experiments and by reasoning with the relations between variables. The created model becomes an artifact students can manipulate to express evolving knowledge. Since modeling-based and simulation-based learning are so closely related, the difference between the two may reveal typical characteristics of modeling knowledge.

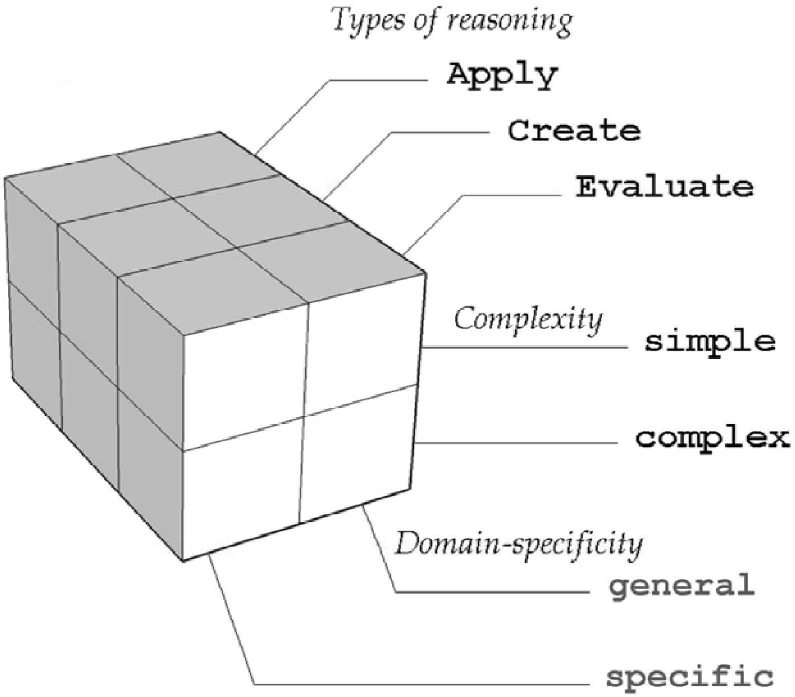


Figure 5-1 The modeling knowledge ACE framework

In a previous study we developed and validated a test about the modeling of dynamic systems based on the ACE framework (see Chapter 3). In a subsequent experimental study, the differences in learning outcomes between modeling-based learning and expository teaching were examined using the domain-specific part of the ACE modeling test as posttest (see Chapter 4). The domain-general ACE subtest was used as a pretest to control for a priori individual differences. The test revealed a difference on the complex items in favor of the modeling group. More specifically, it was found that the modeling group performed significantly better on the complex evaluate items and on the complex reproduce items. Apparently, the modelers were better able to evaluate complex models and to reproduce complex conceptual knowledge. However, no differences were found between the groups on the create scores and on the apply scores. We assumed that the lack of difference was caused by the limited modeling proficiency achievement level in general by the modeling students. This could have been caused by the restricted amount of time that was available for the modeling activity. Another possible reason may be found in the instruction which may have invited students to copy given models instead of creating them from scratch. This may have led them to working superficially with the models. The basic building blocks of modeling appeared to be hard to understand, which is in line with the findings of Cronin and Gonzalez (2007). For example, many students created a model without the key variable (in this case 'temperature'), possibly because the models presented to them in the instruction had a structure without such a key variable. As a result, the students were not able to create models of phenomena that significantly deviated from the phenomena they had practiced with. Therefore, we changed the modeling instruction and put more emphasis on creating models from their basic components.

5.2 Purpose of the study and hypotheses

The aim of the current study was to examine the differences between modeling and simulation-based instruction. Two groups of students worked with either a modeling-based or a simulation-based approach on the energy household of the Earth. It is hypothesized that modeling students will develop more structural knowledge working with the artifact of a model and therefore will be more proficient in assignments about complex structures. In general, we assume that the types of reasoning represent processes of an increasingly higher level and we expect the modelers to be more proficient in the higher-order processes. As opposed to the complex items, we expect that the simple items overall will invoke knowledge that is equally mastered in both modes of learning and will show no differences.

According to the increasing complexity of the types of reasoning from elementary to higher-order, we have the following hypotheses. We expect the learners in the simulation-based mode of instruction to spend more time learning about the definitions of the domain concepts (as opposed to creating a model) and therefore to be more proficient and extensive in *reproducing* domain knowledge. Furthermore, we expect the modeling students to be more proficient in *applying* knowledge of relations, i.e., giving explanations and predictions of complex system's behavior. The modeling students may use their model for concrete step by step relation reasoning in their explanations, while the learners in the simulation-based mode of instruction only receive feedback from the simulation about the end values of an experiment and are offered no insights about the steps in between. Furthermore, because the instruction is more focused on the building blocks of creating models we expect the modeling students to be more proficient in *creating* models at both levels of complexity.

We expect the two groups to perform equally on *evaluating* models and relations, since evaluating models is involved in both experimenting with a model and with a simulation.

5.3 Method

5.3.1 Participants

Participants were eleventh grade students following a science track from two secondary education schools. From the 85 students who started the experiment 76 completed both sessions. Students' age was in the range of 16-19 (M = 16.97, SD = .60); 40.8 % were male and 59.2 % were female.

5.3.2 Materials

5.3.2.1 Co-Lab learning environment

The Co-Lab software provided a learning environment for each of the two conditions to learn about global warming. One environment was designed for modeling inquiry and contained a simulation of the basic energy model of the Earth, a modeling editor to create and simulate models, graphs and tables to evaluate the data produced by the model, and textual information about the domain. A second environment was designed for simulation learning and contained the same tools except for the modeling editor.

Worksheets with assignments helped the participants to understand the factors in global warming. The assignments contained four parts. The first part was about climate models in general. The modelers practiced creating and reasoning about the basic building blocks of modeling. In comparison with the instruction used in the previous chapter, in order to stimulate higher levels of reasoning while modeling and a more profound modeling behavior, students were explicitly made aware of the different functional parts of a model (stock, inflow, and outflow) and were asked to check their models by reasoning about each part of the model, relating the model at hand to previous examples. They were asked to work with a limited set of predefined situations and also to create models for multiple situations and to reason with the created models. The learners in the simulation-based mode of instruction answered questions about the quality and accuracy of using models to make global warming predictions.

The second part discussed the factors albedo and heat capacity, provided information, and included questions about reasoning with the influence of these factors on the temperature on Earth. This was implemented in different ways for the modelers (who created a model to support the reasoning) and the learners in the simulation-based mode of instruction (who used the simulation). For example, an assignment about the influence of the albedo on the equilibrium temperature asked both groups to predict what would happen with the equilibrium temperature if the albedo was high or low respectively. Subsequently, the modelers were asked to investigate their hypotheses with their model and the learners in the simulation-based mode of instruction with the simulation. The third part was about the equilibrium of temperature. The modelers were asked to reason about equilibrium and connect the reasoning steps to the specific parts of their models. The learners in the simulation-based mode of instruction were asked to reason about equilibrium without connecting to an artifact. The fourth part was about evaluating one's understanding of the domain structure. The modelers were asked to compare their own model's behavior with the external model of the simulation of the phenomenon. The learners in the simulation-based mode of instruction were asked to write a report about the factors influencing the temperature on Earth.

5.3.2.2 Pre- and posttest

Based on the ACE framework, we developed the ACE test consisting of a domain-general and a domain-specific part (see Chapter 4). This same test was used in the current study. We used the domain-general part of the ACE test as a pretest and the domain-specific part as a posttest. The pretest assessed domain-general modeling skills and did not require prior knowledge. The posttest was about the domain the students had learned about in the instruction. The pre- and posttest were paper-and-pencil tests which used a causal concept map notation to represent the dynamic models. The pretest consisted of 25 items about a dynamic system in the fantasy topic of the 'Harmony of the spheres'. The posttest consisted of 24 items in the domain of the 'Black sphere' about the energy of the earth. The distribution among the different dimensions of the framework is shown in Table 5-1. Reproducing conceptual knowledge concerned facts and relation in the domain-dependent test and general modeling concepts in the domain-independent subtest. The latter were added to be able to control for prior knowledge on the modeling formalism used, but does not represent the counterpart of the domain-dependent part of the test.

Table 5-1 Distribution of the number of items in pre- and posttest among the framework dimensions

Number of items	Pretest		Posttest	
	Harmony		Black sphere	
	Simple	complex	simple	complex
Reproduce items*	3	3	3	3
Apply items	3	4	3	3
Create items	2	4	3	3
Evaluate items	3	3	3	3
	11	14	12	12
Total	25		24	

* The test items about the modeling formalism are classified under "reproduce harmony", although this is technically incorrect.

5.3.3 Procedure

The experiment consisted of two sessions of approximately 200 minutes each. The first session was the same for all participants and included an introduction to dynamic systems and to the computer environment that lasted 50 minutes. All participants worked through a technical manual of the model editor for 50 minutes and conducted exercises with a given model of a leaking bucket for again 50 minutes. In the last 50 minutes of the first session the pretest was administered. For the second session, the participants were divided into two conditions based on the pretest scores such that student groups in both conditions had a similar mean and standard deviation on the pretest. Each combination of school, teacher, class, and gender was included in both conditions. The second session took place three weeks after the first session. The modeling condition worked on a model about the energy of the earth in an environment with a simulation of the phenomenon and with a model editor in which they could create runnable models. The simulation condition worked on similar assignments about the energy of the earth using only the simulation. Finally, in approximately 50 minutes the black sphere posttest was administered.

5.3.4 Analysis

The test responses were scored using a coding scheme with an interrater reliability of .77 (see Chapter 3). We analyzed the sum scores of the two experimental groups with an analysis of variance with the pretest score as a covariate. Pretest *subscores* were used as a covariate in the analysis of black sphere *subscores*. For example, in the analysis of the scores on the simple apply subtest we used the scores on the pretest simple apply subtest as a covariate. The exception is conceptual knowledge for which no domain-independent counterpart exists and therefore no covariate was used.

In addition to the analysis of sum scores, we analyzed the occurrence of specific content elements in the item responses. The content elements were analyzed with respect to domain knowledge and create errors respectively and were counted in both simple and complex items, because we were interested in an indication of the overall prevalence of the elements.

5.4 Results

Due to the experimental design (see under Procedure) both conditions had equivalent scores on the pretest. There were no significant differences on the pretest total score and on the scores for the subscales (see Table 5-2).

We compared the mean black sphere (sub) scores of the two conditions using analyses of variance (see Table 5-3). In order to account for individual differences on prior modeling skills, we used the pretest (sub)score as a covariate. As expected we found no significant difference between the conditions on the total black sphere score. Also, we did not find a significant difference on the simple items overall. In line with our expectations, we found that the modelers performed significantly better on the complex items overall ($F(1, 74) = 7.036$, $p = .010$, partial $\eta^2 = .088$).

We found significant differences on the subscales of the ACE black sphere test. As hypothesized, the modelers were more proficient in *applying knowledge* of relations and performed significantly better on the apply items ($F(1, 74) = 7.949$, $p = .006$, partial $\eta^2 = .098$). More specifically, a difference was found for explaining and predicting system's behavior on a complex level, i.e., for the complex apply items ($F(1, 74) = 19.619$, $p = .000$, partial $\eta^2 = .212$).

As expected, the modelers were more proficient in *creating models*. A significant difference was found in favor of the modelers for the simple items ($F(1, 74) = 6.679$, $p = .012$, partial $\eta^2 = .084$). No significant difference was found for the complex items. As our modified instruction put more focus on the basic elements of models, this did not come unexpected.

Table 5-2 Means and standard deviations of the pretest (sub)scores for the two conditions

		Harmony pretest				
		Simulation (*n* = 39)		Modeling (*n* = 37)		max
Overall	simple	5.54	(1.67)	5.27	(1.49)	11
	complex	6.36	(2.55)	6.29	(2.50)	14
	total	11.91	(3.97)	11.56	(3.32)	25
Reproduce	simple	0.54	(0.76)	0.52	(0.61)	3
	complex	1.28	(0.45)	1.36	(0.50)	3
	total	1.81	(0.96)	1.89	(0.90)	6
Apply	simple	1.34	(0.55)	1.20	(0.58)	3
	complex	2.25	(1.14)	1.89	(1.12)	4
	total	3.60	(1.48)	3.09	(1.37)	7
Create	simple	1.68	(0.58)	1.64	(0.64)	3
	complex	1.49	(1.05)	1.62	(1.10)	3
	total	3.16	(1.44)	3.26	(1.51)	6
Evaluate	simple	1.99	(0.67)	1.91	(0.83)	2
	complex	1.35	(0.85)	1.42	(0.88)	4
	total	3.33	(1.30)	3.32	(1.24)	6

A further analysis of the students' answers with respect to create errors revealed an overall significant difference between the conditions. We counted the following errors: non-specific variable name, non-relevant variable name, error in using an extra variable (outside of the context), drawing a direct relation that is indirectly already present, and drawing an alternative representation. In both simple and complex create items the modelers made significantly less errors than the learners in the simulation-based mode of instruction ($F(1, 74) = 8.687$, $p = .004$, partial $\eta^2 = .105$), indicating that their modeling ability indeed increased.

As expected, no difference was found on the *evaluate* items. Furthermore, as hypothesized, the learners in the simulation-based mode of instruction gained more domain knowledge and performed better on *reproducing conceptual knowledge*, however, only on the simple items ($F(1, 74) = 11.799$, $p = .001$, partial $\eta^2 = .138$).

Table 5-3 Means and standard deviations of the posttest black sphere (sub)scores for the two conditions

		Black sphere posttest				
		Simulation (n = 39)		Modeling (n = 37)		max
Overall	simple	6.45	(2.10)	6.07	(1.72)	12
	complex	4.71*	(2.13)	5.72*	(1.97)	12
	total	11.16	(3.78)	11.79	(3.31)	24
Reproduce	simple	2.05	(0.97)	1.29*	(0.97)	3
	complex	1.60	(0.82)	1.79	(0.86)	3
	total	3.65	(1.37)	3.08	(1.43)	6
Apply	simple	1.26	(0.80)	1.26	(0.76)	3
	complex	1.07*	(0.77)	1.64*	(0.76)	3
	total	2.33*	(1.33)	2.90*	(1.30)	6
Create	simple	1.71*	(0.74)	2.06*	(0.42)	3
	complex	1.24	(0.73)	1.38	(0.59)	3
	total	2.95	(1.36)	3.43	(0.88)	6
Evaluate	simple	1.43	(0.53)	1.46	(0.58)	3
	complex	0.80	(0.61)	0.92	(0.66)	3
	total	2.23	(0.85)	2.38	(0.96)	6

* Means differ at $p < .05$ in the analysis of variance.

To get an impression of the extent to which students referred to domain knowledge, we counted the occurrences of referring to a definition (albedo, heat capacity), the reasoning with 'shortcuts' (as opposed to step by step reasoning), and the use of extra variables in creating a model. Over all items other than the items on reproducing conceptual knowledge, the simulation condition used this type of prior domain knowledge significantly more than the modeling condition ($F(1, 74) = 4.485$, $p = .038$, partial $\eta^2 = .057$) (see Table 5-4).

Table 5-4 Number of references to domain knowledge and number of create errors in the posttest black sphere items for the two conditions

	Black sphere posttest				
	Simulation (*n* = 39)		Modeling (*n* = 37)		max
Prior knowledge	3.62*	(1.87)	2.70*	(1.88)	23
Create errors	2.15*	(0.96)	1.51*	(0.93)	15

* Means differ at $p < .05$ in the analysis of variance.

5.5 Discussion

The current study extends our understanding of the specific learning gains of computer modeling by using a test about modeling in a comparative study with two modes of instruction. We compared two groups of students in a learning activity about the energy of the earth, one using a computer environment including both a simulation and a model editor for creating runnable models, and one group with an environment including only the simulation. We used the modeling ACE test to reveal the specific differences and similarities between the two groups of learners. We expected to find differences in favor of the modeling students for the complex items and, more specifically, for the apply and the create types of reasoning. Differences in favor of the simulation-only students were expected for the reproduce type of reasoning. We expected no differences in learning outcomes for the evaluate type of reasoning since both groups equally were performing experiments, with a self-made model and with the simulation respectively.

The results show that our hypotheses were confirmed. It was found that the modelers performed significantly better on the complex items overall. These results are in line with the previous comparative study described in Chapter 4 in which modelers performed significantly better on complex items compared to students in an expository mode of instruction. With respect to the types of reasoning, significant differences were found for reproduce, apply, and create and these differences occurred on only one particular level of complexity. The learners in the simulation-based mode of instruction were significantly better reproducers of *simple* knowledge. Moreover, the learners in the simulation-based mode of instruction used significantly more prior knowledge in the items overall. The modelers were significantly better in applying *complex* knowledge of a model and in *simple* create items. No differences were found for the complex create items. As the modified instruction focused more on the basics of modeling this is not unexpected. As expected, no differences were found for evaluate items, neither simple nor complex.

It was hypothesized that the learners in the simulation-based instruction would be more proficient in reproducing knowledge. As described above, only for the simple reproduce items a difference was found. An explanation of the lack of difference in performance on the *complex* reproduce items may be that reproducing complex conceptual knowledge to a larger extent relates to more complex reasoning. As the modelers perform better on the complex apply items, they are more likely to use this reasoning ability in the complex reproduce items, concealing a possible smaller learning gain than the learners in the simulation-based mode of instruction. The simple apply items are equally well performed by both groups, suggesting that these items cover an ability acquired in both learning conditions.

With respect to reproducing conceptual knowledge, a difference arose between the current and the previous study. In the previous study (see Chapter 4) we compared modeling with expository instruction. We expected that the learners in the expository mode of instruction would perform better on reproducing knowledge since they spent more time on working with the definitions of the domain concepts. Nonetheless, in that study the modelers performed significantly better on the complex reproduce items than the learners in the expository mode of instruction. However, in the present study we found a difference on reproducing simple conceptual knowledge between the modelers and the learners in the simulation-based mode of instruction in favor of the latter. Moreover, the learners in the simulation-based mode of instruction used significantly more prior knowledge elements in the items other than the reproduce items. This raises the question what caused the contrast in outcomes between the expository and simulation-based mode of instruction. Learners in both expository and simulation-based modes of instruction completed the learning activity by writing a report about what they had learned in the energy of the earth assignments. An explanation may be that the learners in the simulation-based mode of instruction worked with the concepts and definitions in a more engaging activity in which they received feedback from the simulation and therefore the factual knowledge was better understood and easier to recall.

It must be noted that in both the previous (see Chapter 4) and the current study the evaluate scores on the complex items were rather low. In the previous study the modelers scored significantly better than the learners in the expository mode of instruction (though still low) on these particular items, but in the current study there were no differences between the two conditions. This comparative result is logical, but the low scores indicate that the complex evaluate items are rather difficult for the students. One explanation of the low scores may be that the test's scoring method by design is aiming at a slightly higher level to avoid losing information in a 'ceiling effect'. Another explanation may be that the learners in both modeling-based and simulation-based modes of instruction are implicitly learning about evaluating models and relations, but that they are not able to make the knowledge explicit on the test. A possible solution to improve this explicitness is to add cues to the instruction to make students aware of the steps in evaluating.

A general pattern that arose from this study is the difference in 'instructional benefit' with respect to simple and complex items and processes. The types of reasoning represent processes that are increasingly more demanding, from the straightforward *reproducing* conceptual knowledge to the higher-order processes of *applying*, *creating*, and *evaluating*. The findings of this study indicate a benefit for the learners in the simulation-based instruction on the simplest process of reproducing conceptual knowledge. Progressively, the modelers benefited more on the complex items and the more demanding processes. However, neither of both groups reached a high level for the process of evaluating.

We can draw conclusions both related to the test and to the instruction. First, the test reveals significant differences within the subscales of the types of reasoning and thereby appears to be sensitive enough to detect significant differences between modeling-based and simulation-based learning on these specific cognitive areas. Finding differences on particular levels of complexity, suggests that the complexity dimension makes an effective distinction within the dimension of the types of reasoning. Second, the modeling-based instruction seems to enhance measurable learning gains. The knowledge gain in the apply category shows that the modelers are able to reason with their model and to give explanations of complex relations within a model. The modelers succeeded to achieve well-reasoned effective modeling behavior, albeit for the creation of models at the simple level. The effect of stepwise practice of the creation and meaning of a model seems to be useful and seems to cause fine-grained differences in model-based reasoning skills.

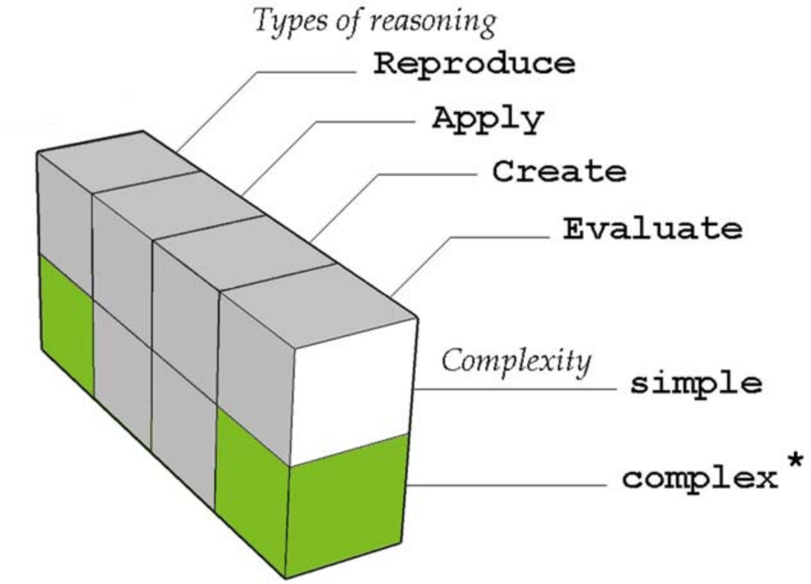
Chapter 6 Combined analysis

Abstract

In this chapter a combined analysis is presented regarding the data of the two comparative studies described in Chapter 4 and 5. Besides the pairwise comparisons that have been performed in each comparative study, a combined analysis would allow for a comparison of conditions across the studies, for example the modeling students in the first comparative study and the modelers in the revised instruction. The results of this combined analysis show that the revised modeling instruction had a positive effect on the performance for *applying*, but not for *creating*. Additionally, a more detailed analysis of content elements suggests that the strategy to use domain knowledge might hinder the application of reasoning skills. Altogether, the results indicate that different modes of instruction enhance different types of knowledge. A general pattern arises with respect to simple and complex items and types of reasoning: the more complex the learning outcomes, the more benefit a learner experiences from modeling-based instruction.

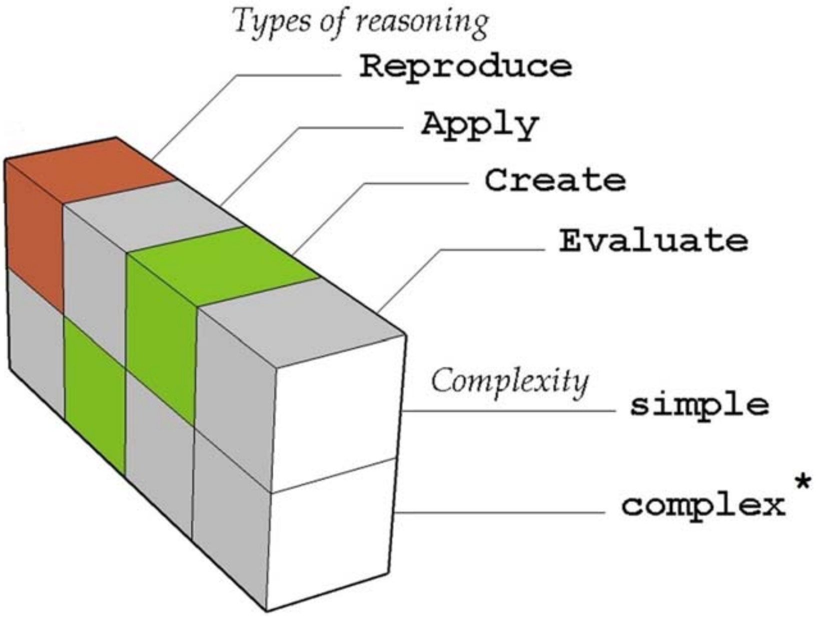
6.1 Introduction

The studies presented in the two previous chapters provide information about the comparison of instruction based on computer modeling with other modes of instruction. In the first comparative study described in Chapter 4, we compared modeling with expository instruction. This comparison revealed differences with regard to the processes of reproducing and evaluating for complex items (see Figure 6-1). Against our expectations, no differences were found for the core modeling processes of applying and creating. In the second comparative study described in Chapter 5, we revised the modeling instruction to put more emphasis on reasoning with the building blocks of modeling and compared modeling with simulation-based instruction. The results of the second study showed differences with regard to the processes of reproducing, apply, and create, and, as expected, not with respect to evaluating (Figure 6-2).



Note. Green indicates that the modeling condition performed significantly better than the expository condition.
* The modeling condition performed significantly better on the complex items overall.

Figure 6-1 Differences in the domain-specific posttest subscores between the expository and the modeling conditions in the first comparative study



Note. Red indicates that the modeling condition performed significantly lower than the simulation condition; green indicates that the modeling condition performed significantly better.
* The modeling condition performed significantly better on the complex items overall.

Figure 6-2 Differences in the domain-specific posttest subscores between the simulation and the modeling conditions in the second comparative study

In total four modes of instruction have been examined, among which two versions of modeling instruction. Apart from the two pairwise comparisons we have presented in the earlier chapters, a combined analysis of the data from the two comparative studies would also allow for the examination of differences across studies. For instance, it is of interest to examine whether our modification of the modeling instruction indeed led to an improvement. Also, comparing simulation-based instruction with expository instruction is interesting with respect to scientific reasoning skills.

In this chapter we present such a combined analysis of the response data of the two comparative studies together (n = 150). We could do this after having checked that the groups were indeed comparable The two comparative studies were conducted at the same two schools, at the same period of the school year. The students were following the same physics classes taught by the same teachers, except for 24 out of 76 students in the second comparative study who had a different teacher. Furthermore, all students completed the same pretest allowing us to check for any differences between groups. Because the two studies have been conducted in these comparable circumstances, we could combine the data and perform an analysis of variance of the sum scores of the conditions in both studies. We will analyze the differences with respect to the simple and complex parts of the subscales of the ACE framework and additionally of reproducing conceptual knowledge. To gain insight into strategies used in the responses a more detailed analysis of specific content elements in the responses were analyzed.

6.2 Research question

The findings of the two comparative studies each show differences and similarities between two experimental conditions (see Figures 6-1 and 6-2). An analysis of the combined data may reveal differences and similarities comparing the conditions of the different studies. Our research questions is:

How do all four conditions over both studies perform in relation to each other on the domain-specific simple and complex parts of the subscales of the ACE framework?

And, more in particular:

Can effects be found of the revision of the modeling instruction in the second comparative study with respect to the apply and create items?

6.3 Results of combined analysis of experimental data

The pretest scores for the four experimental groups in the two comparative studies did not differ significantly (see Table 6-1).

Table 6-1 Means and standard deviations of the pretest scores of the experimental groups

	Condition							
	Expository		Modeling 1		Simulation		Modeling 2	
	(*n* = 36)		(*n* = 38)		(*n* = 39)		(*n* = 37)	
Pretest score	11.01	(3.29)	11.09	(3.67)	11.91	(3.98)	11.56	(3.32)

Table 6-2 Means and standard deviations of the black sphere (sub)scores for the types of reasoning of the experimental groups

	Condition							
	Expository		Modeling 1		Simulation		Modeling 2	
	(n = 36)		(n = 38)		(n = 39)		(n = 37)	
Overall								
simple	6.59	(1.38)	6.58	(1.68)	6.45	(2.10)	6.07	(1.72)
complex	3.67_a	(1.50)	4.72	(1.72)	4.71	(2.13)	5.72_b	(1.97)
total	10.26	(2.48)	11.30	(3.10)	11.16	(3.78)	11.79	(3.31)
Reproduce								
simple	2.00_a	(0.80)	1.74	(0.84)	2.05_a	(0.97)	1.29_b	(0.97)
complex	1.07_a	(0.71)	1.50	(0.69)	1.60_b	(0.82)	1.79_b	(0.86)
total	3.07	(1.09)	3.23	(1.28)	3.65	(1.37)	3.08	(1.43)
Apply								
simple	1.05	(0.69)	1.22	(0.70)	1.26	(0.80)	1.26	(0.76)
complex	0.90_a	(0.69)	1.12_a	(0.71)	1.07_a	(0.78)	1.64_b	(0.76)
total	1.95_a	(1.16)	2.33	(1.18)	2.33	(1.33)	2.90_b	(1.30)
Create								
simple	2.09	(0.69)	2.07	(0.85)	1.71	(0.74)	2.06	(0.42)
complex	1.16	(0.74)	1.26	(0.76)	1.24	(0.73)	1.38	(0.59)
total	3.24	(1.34)	3.33	(1.48)	2.95	(1.36)	3.44	(0.88)
Evaluate								
simple	1.46	(0.59)	1.56	(0.66)	1.43	(0.53)	1.46	(0.58)
complex	0.54	(0.53)	0.85	(0.63)	0.80	(0.61)	0.92	(0.66)
total	2.00	(0.79)	2.41	(0.99)	2.23	(0.85)	2.38	(0.96)

Note. Means in the same row with different subscripts differ at $p < .05$ in the Scheffé post-hoc test.

6.3.1 Comparing performance for the types of reasoning and complexity for the black sphere domain

We performed an analysis of variance for the sum scores on the black sphere subtests for each of the types of reasoning and complexity with the pretest subscore as a covariate. Using the Scheffé post hoc criterion for significance the results show significant differences for the complex items overall and for reproduce simple, reproduce complex, and apply complex (see Table 6-2). The learners in the revised modeling-based instruction ('modeling 2') performed significantly better on the complex items overall than the learners in the expository mode of instruction. The learners in the revised modeling-based instruction performed significantly worse on the reproduce simple subtest than in the expository and simulation mode of instruction. The learners in the expository mode of instruction performed significantly worse on the reproduce complex subtest than in the revised modeling-based ('modeling 2') and simulation mode of instruction. For the apply complex subtest the intergroup comparisons indicated that the modelers in the second study perform significantly better than the students in the three other conditions.

6.3.2 Response content analysis

An analysis of variance of the amount of prior domain knowledge used in the students' answers showed that the learners in the simulation-based mode of instruction significantly used more domain knowledge than in the two conditions in the first study. The learners in the simulation-based mode of instruction made significantly more create errors than the modeling students in the first study (see Table 6-3).

Table 6-3 Number of references to domain knowledge and number of create errors in the posttest black sphere items for the two conditions

	Condition							
	Expository		Modeling 1		Simulation		Modeling 2	
	(*n* = 36)		(*n* = 38)		(*n* = 39)		(*n* = 37)	
Domain knowledge	2.44_a	(1.38)	2.00_a	(1.49)	3.62_b	(1.87)	2.70	(1.88)
Create errors	1.67	(1.29)	1.34_a	(1.24)	2.15_b	(0.96)	1.51	(0.93)

Note. Means in the same row with different subscripts differ at $p < .05$ in the Scheffé post-hoc test.

6.4 Conclusion

The analysis of the combined data of the two comparative studies presented in Chapters 4 and 5 showed differences between conditions we could not compare before and provided us with additional insight into the effects of the different modes of instruction.

Differences were found for the complex items overall and for several parts of the subscales for the types of reasoning. First, for the complex items overall the modelers in the revised condition performed better than the learners in the expository mode of instruction. No differences were found between other pairs of conditions. Though the pairwise comparisons in both comparative studies showed significant differences in favor of the modeling condition, these differences were not large enough to remain significant in the combined analysis with four groups. The Scheffé post hoc test corrects for the type I error, the probability of a false rejection of the null hypothesis regarding all of the six comparisons being made and therefore requires a lower significance level for the single comparisons (Sato, 1996). Nevertheless, the results for the complex items overall suggest the trend that the modes of expository, simulation-based, and modeling-based instruction progressively enhance the acquisition of reasoning skills for complex situations.

Second, the expository condition performed significantly better on *reproducing simple* conceptual knowledge than the modeling condition in the second comparative study. This finding suggests that simple conceptual knowledge is best acquired without the task of creating a model in an expository or a simulation-based mode of instruction. Third, the modeling condition in the second study performed significantly better on *applying complex* knowledge than all of the other conditions. This finding suggests that the revised instruction seems successful in focusing on profound and well-reasoned modeling behavior resulting in higher applying reasoning skills. Fourth, the expository condition performed significantly worse on the *complex evaluate* subscale than the revised modeling condition. The mean scores of the other two conditions suggest a trend that complex evaluate skills are acquired equally well in both a modeling-based and a simulation-based mode of instruction. Fifth, with respect to *creating simple* models, no significant differences were found. The pairwise comparison of the simulation condition and the modeling condition in the second study showed significant differences in favor of the modeling condition, but this difference was not large enough to remain significant in the combined analysis with four groups.

The more detailed analysis of the learners' responses indicates that the simulation condition in the revised instruction used more domain knowledge in their answers and made more create errors than the other conditions. The pairwise comparison of the modes of instruction in the second comparative study showed significant differences with respect to the frequency of the use of domain knowledge and the number of create errors. These differences were not large enough to remain significant in the combined analysis. Nevertheless, the differences suggest that using domain knowledge might be a strategy that hinders the application of reasoning skills.

With respect to the more specific second question about the effect of the revised instruction, a salient effect was found for the complex apply test. The modelers who receive the revised instruction performed significantly better than the other conditions. However, the revised instruction did not appear to have an effect on the create performance.

In conclusion, different modes of instruction seem to enhance different types of knowledge. The acquisition of simple conceptual knowledge seems to be best realized in a mode of instruction that does not involve the construction of a model. Applying complex knowledge seems to be best mastered in a modeling-based mode of instruction. Evaluating complex data seems to be best mastered in a modeling-based mode of instruction, but might equally well be learned in a simulation-based mode of instruction. A general pattern arises with respect to simple and complex items and types of reasoning. The types of reasoning are increasingly more demanding, from the straightforward *reproducing* conceptual knowledge to the higher-order processes of *applying*, *creating*, and *evaluating*. The more complex the learning outcomes, the more benefit a learner experiences from modeling-based instruction.

Chapter 7 Discussion and conclusions

7.1 Introduction

In the introductory chapter to this dissertation, we argued that the ongoing debates about the effectiveness of new modes of instruction would be well-served by a theoretically grounded and evidence-based view of the learning outcomes that can be expected from such modes of instruction. In this dissertation we focused specifically on the learning outcomes of computer modeling. Computer modeling has as a defining characteristic that students construct and/or modify executable computer models of dynamic phenomena (following Löhner, 2005). The main research question in this dissertation was:

What specific learning outcomes can be expected from computer modeling and how can they be measured?

To answer the research question, we developed the ACE framework, which was described in Chapter 2. This framework distinguishes learning outcomes on three dimensions: 'type of reasoning', 'complexity', and 'domain-specificity'. This framework was used as a basis for the creation of the ACE test, which measures learners' knowledge along these three dimensions. We created the ACE test for the domain of global warming and a parallel 'fantasy domain' and subsequently validated this test in three empirical studies. The first of these studies explored construct validity, whereas the discriminative validity of the test was investigated in the other two studies.

7.2 Construction of the ACE test

The three dimensions of the ACE framework describe kinds and levels of learners' knowledge. The 'type of reasoning' dimension takes values that describe the processes of reasoning: *applying* concepts and relations to given situations, *creating* new variables and relations in a model, and *evaluating* a model. Complexity refers to whether learners are able to exert these processes in either complex or simple situations. Finally, domain-specificity concerns whether the knowledge is bound to a specific domain or is domain-independent.

The ACE framework results in a 3x2x2 matrix, with cells that represent processes of a given type that are either domain-specific or domain-independent and either simple or complex. In order to use the framework to create a test that can measure knowledge along these dimensions, items must be created for each of the resulting twelve cells. We did so for the domain of global warming by constructing Apply, Create and Evaluate items for single variables and relations in the domain as well as for more complex substructures of a model of the domain. All of these items were domain-specific. In order to create domain-independent items, we developed a parallel fantasy domain with the same conceptual structure as the global warming domain, but with no relation to any reality.

We also created domain-specific items at both simple and complex levels for *reproducing conceptual knowledge*. By its nature, reproducing knowledge in a domain-independent way is impossible. This yields two additional cells on top of the twelve described above. Two more cells were added to the test by including items on factual knowledge about the modeling formalism, again at two levels: simple and complex.

7.3 Validation of the ACE test

The validity of the ACE test was examined in an initial empirical study among 131 participants with varying levels of modeling experience and background knowledge. We collected test responses and analyzed the responses to find evidence for the existence of these dimensions in the response data. The findings of this study confirmed that the composite skill of model-based reasoning can be decomposed into four unidimensional subscales, namely, Apply, Create, Evaluate, and Reproduce, thus confirming the validity of the main dimension of 'type of reasoning' of the ACE framework.

With respect to the dimension of complexity, we found that the items about complex situations were significantly more difficult than the items about simple situations. This finding confirmed the validity of the dimension of complexity. The validation of the third dimension, domain-specificity, falls beyond the scope of this dissertation work, because generalizing domain-specific reasoning skills to a domain-general context is particularly likely to occur after much more substantial modeling experience.

Next, we performed two comparative studies to contrast the learning outcomes of different modes of instruction. In the first study, we compared computer modeling with expository instruction. In the second comparative study, modeling was compared with simulation-based inquiry instruction. In this second study, we used a revised version of the modeling condition that emphasized reasoning with the basic building blocks of a model. In both studies, prior modeling knowledge was measured using the domain-independent section of the ACE test, while learning outcomes were assessed using the domain-specific subtest of the ACE test about global warming. By comparing modeling with a mode of instruction offering quite a contrast in the first study and with a mode that had more in common with modeling in the second study, the ACE test was 'put to the test' in a gradually more rigorous way.

In both studies, the modelers scored significantly higher on the complex items, summed over all reasoning types, than the students in the alternative mode of instruction (expository and simulation-based instruction respectively). This salient benefit for the modeling condition on the complex part of the subscales suggests that modeling contributes especially to reasoning with complex structures. Moreover, within the dimension of complexity specific differences between conditions were found for the reasoning types. In the first comparative study, the modelers performed significantly better on the *complex* evaluate and *complex* reproduce items than the students in the expository mode of instruction. In the complex evaluate items, the modelers showed a greater ability to judge the differences and similarities of complex models and to judge conclusions based on experimental data. In the complex reproduce items, the modelers were better able to reproduce the concepts in the domain in relation to each other.

In the second comparative study, a different picture emerged: the modelers performed significantly better on the *complex* apply and *simple* create items than the students in the simulation-based mode of instruction and significantly worse on the *simple* reproduce items. This finding suggests that students who explicitly practiced reasoning with the different basic building blocks of a model (stock, inflow, and outflow) performed significantly better on applying knowledge of complex model structures and on creating simple model structures. Moreover, for the complex apply items the modeling students in the revised mode of instruction performed better than the modelers from the first study who received instruction without the explicit reasoning practice. No difference was found when comparing the performance of these two modeling groups on the simple create items. It seems that the revision of instruction affects higher-order reasoning skills, but does not induce create abilities. It can be noted that the create items asked for modeling of a phenomenon that was slightly different from the models in the instruction. It might be possible that this required a level of ability that had not yet been attained.

The differences and similarities observed provide evidence of the discriminative validity of the ACE test. Finding differences on particular levels of complexity within a subscale suggests that the complexity dimension makes an effective distinction and that the ACE test is able to detect these differences. In the case of a *lack* of difference the question is whether the test items are not sensitive enough to reveal differences on a specific part of the test or whether there were no differences. Because we expected benefits of modeling mainly for items about complex structures, we conclude that the simple items overall invoke knowledge that can be equally well acquired in the expository, modeling-based, and simulation-based modes of learning.

A more detailed analysis of the answer elements with respect to the use of domain knowledge revealed a possible relation between references to domain knowledge and the acquired level of modeling proficiency. In the first comparative study, the students in the expository mode of instruction frequently used domain knowledge as expected, but unexpectedly, also the modelers used domain knowledge frequently. In the second study, the learners in the simulation-based mode of instruction expressed significantly more domain knowledge in the item responses overall than the modeling-based learners. This result might suggest that there is a higher tendency to use domain knowledge in cases of lower reasoning abilities. The test requires the students to use their model-based reasoning skills. Apparently, when students are able to apply these skills, there is less urge to use domain knowledge. Conversely, students lacking model-based reasoning skills more often try to fall back on their existing domain knowledge.

7.4 Effects of instruction

Apart from validating the test, the studies presented in this dissertation contribute to insight into the effects of the different modes of instruction. The modeling-based, simulation-based, and expository modes of instruction each affected specific learning outcomes that we distinguish in the ACE framework. Specific kinds of instruction appear to be related to learning outcomes as measured by specific parts of the ACE test. This would mean that specific modes of instruction might be used to stimulate the development of specific reasoning skills. After having examined the differences and similarities in performance in each comparative study, we now try to distill an overall picture from the results.

A general pattern that arose from the two comparative studies is the difference in performance between the groups with respect to simple and complex items and processes. Overall, the students in the modeling-based mode of instruction performed better on the complex items. With regard to the types of reasoning, the processes are increasingly more demanding, from the straightforward *reproducing* conceptual knowledge to the higher-order processes of *applying*, *creating*, and *evaluating*. The findings indicate a benefit for the learners in the expository and simulation-based instruction on the simplest process of reproducing conceptual knowledge. The modelers in the second comparative study benefited progressively more on the complex items and the more demanding processes of applying and creating. However, neither group reached a high level for the process of evaluating in this study.

Based on the combined analysis of the data from both comparative studies, presented in Chapter 6, recommendations can be made for the most effective mode of instruction with respect to a particular learning effect. First, in order to learn simple conceptual domain knowledge, creating a model seems to be disadvantageous. The best performance appears to be realized in an expository or a simulation-based mode of instruction.

Second, in order to learn reasoning with complex structures, modeling seems to be most effective. Students using the second version of the modeling instruction were superior on complex apply items compared to all other conditions.

Third, with respect to create abilities no differences were found between the conditions. In the revised instruction the modeling-based learners performed better than the learners in the simulation-based mode of instruction on the simple create items. However, when comparing the performance of the modelers in the second and first study, there is no significant difference. It was assumed that additional practice in reasoning with basic model building blocks would improve create abilities, but this appeared not to be the case. This suggests that practicing reasoning with parts of a model does not automatically lead to the acquisition of basic create abilities and well-reasoned modeling behavior.

Finally, evaluating skills appear to be equally well acquired in both simulation-based and modeling-based modes of instruction. To develop these skills it does not seem to make a difference whether experiments are performed using a simulation or a self-created model. An expository mode of instruction appears to be less effective.

7.5 Competing frameworks

As we began our research, there were virtually no studies providing systematic descriptions of modeling knowledge. During the time frame of our studies other frameworks have been developed that specifically focus on modeling competences or inquiry reasoning skills in science. The multitude of studies illustrates the relevance of the topic and the more general need for an assessment approach to modeling skills. A description of these frameworks is given in the following paragraphs. The fact that other researchers identified similar aspects of modeling learning outcomes, especially the main dimension of the types of reasoning, provides additional justification for the choices made in composing our modeling knowledge framework.

In order to facilitate the development of assessment instruments, Kauertz and Fischer (2006) constructed a framework for the content structure of physics. Their framework describes multiple levels of complexity in order to be able to assess students' scientific competence in the context of physics for students ranging from grades 7 to 10. From this perspective, their framework contains a hierarchical structure describing six levels of complexity and three kinds of cognitive activities. The cognitive activities are *remembering*, *structuring*, and *exploring*. By investigating a test based on their framework, Kauertz and Fischer found evidence that levels of complexity defined in the model corresponded with item difficulty. Though not specifically designed for modeling, Kauertz and Fischer's cognitive activities bear strong resemblance with the ACE types of reasoning. However, although they also define a dimension of complexity, their dimension of complexity is directly related to item difficulty, reflecting their consideration of complexity as a gradual scale, whereas the ACE framework considers simple and complex reasoning as qualitatively different processes.

Maaß (2006) described competencies in the context of mathematical modeling. The basic competencies related to the process of modeling in her framework are divided in five groups of subcompetencies: first, *understanding the real problem and creating a model based on reality*; second, *setting up a mathematical model from the real model*; third, *solving mathematical questions within this mathematical model*; fourth, *interpreting mathematical results in a real situation*; and, fifth, *validating the solution*. The first two of these groups of competencies form a more detailed account of our 'create' process, including a strong emphasis on mathematical skills. In the context of Maaß' framework, validating the solution is more directed at mathematical structure rather than evaluating by reasoning with empirical data, as is the case in our context.

Schecker and Einhaus (2007) developed a framework for the systematic description of science competencies. Their Bremen-Oldenburger Kompetenzmodell for science competence (BOlKo) comprises the following five dimensions: *content area or basic concepts*, *processes*, *context*, *expertise*, and *cognitive demand*. The dimension of 'content area or basic concepts' defines the subject matter in which the science competence is to be applied. The dimension of 'processes' defines the different ways in which science competencies can be applied, for example use factual and conceptual knowledge, apply methods in experimentation and modeling, communicate, and judge. The dimension of 'context' defines the framing of the modeling problem, for example a didactic situation within a school subject or a problem posed in a professional environment. The dimension of 'expertise' reflects the type of demand on science competence, i.e., can the problem be solved with everyday life expertise, by reproducing content knowledge, by applying knowledge in a familiar situation, or by transferring to a new situation. The dimension of 'cognitive demand' defines the two components of analytical thinking and creativity or productivity. In an empirical study with 26 experts and 680 students from grades 9 to 13, Schecker and Einhaus investigated the validity of a test based on their framework; one of their conclusions is that the framework needs a complexity dimension that differentiates in a more detailed way between students with different levels of science competencies. The BOlKo framework with its five dimensions is less focused on students' reasoning processes and gives more attention to the educational context of the science curriculum. This makes it difficult to compare with the ACE framework, although certainly similarities exist, for example, in the dimension of 'expertise'.

Lopes and Costa (2007) perceive modeling as a process for the construction and use of science concepts that mobilize various specific competences. They regard modeling in a general sense and not as the computer modeling of dynamic systems specifically. In their framework for evaluating competencies with respect to this general definition of modeling they identified the following dimensions: *way of facing*, *conceptualization*, and *operative work*. The dimension of 'way of facing' addresses the use of certain properties or relations of the concepts and analysis of the problem-situation. The dimension of 'conceptualization' applies to objects and events in a specific context and whether objects and events are coherent and suitable for the context. The dimension of 'operative work' includes dealing with relations between the variables, types of relations, physical quantities, predictive capacity, clarity of certain operations (for example the choice of a coordinate system), and symbolic language. Although the dimensions reflect a more general view on modeling competencies, they describe processes related to model-based reasoning, such as defining the concepts to be used, predicting, explaining, and evaluating relations. Though posed at a more general level, these processes are similar to our processes of applying, creating, and evaluating. In proposing their framework, Lopes and Costa stress the importance of creating a methodology to systematically evaluate modeling competences. Lopes and Costa validated the framework in a study with a diverse group of students ranging from 9th graders to PhD students and concluded that despite the complexity of the involved modeling competencies involved, it is possible to evaluate the various types of information using only written information.

The competing frameworks we discussed in this section provide different viewpoints on the competencies associated with modeling. Most of these frameworks are designed for broader contexts than computer modeling.

On the flip side of this broader view is that they tend to describe reasoning processes less precisely.

7.6 Implications for practice and future research

The ACE framework developed in this research provides a classification of learning outcomes of computer modeling that appears to be useful for developing a practical test instrument. The ACE test is able to reveal differential effects for the specific model-based reasoning skills that can be acquired by instruction based on computer modeling. Thereby, the ACE framework implements a feasible path from theory about model-based reasoning to a practical operationalization of the main model-based reasoning processes into test items.

Although the ACE test that we presented in this dissertation was developed for a specific domain, the ACE framework enables the development of test items in other domains. The development of test items for our ACE test exemplifies how this can be done in other domains as well. This means that by applying similar item formats, parallel test item sets can be constructed. For the type of reasoning 'apply', one of the item construction formats involved mentally simulating the effect of two different values of a variable on another variable connected by means of two intermediary variables. This construction format is easily applied in another domain that has an appropriate domain structure containing variables that are connected in two intermediate steps.

Application of the ACE framework to develop a test in another domain requires the following steps. First, a model of the target domain must be available or must be developed by a domain expert. This model must provide a sufficient number of variables and relations as well as sufficiently complex structures to allow for distinguishing levels of complexity. Second, a parallel domain-general model must be developed. This can be done by replacing all variables in the domain model by abstract names and developing a ‘cover story’ about a fantasy subject in which these variables play a role, as was done in developing the harmony of the spheres model. Finally, the test items must be constructed for all cells in the ACE framework, for which our test can serve as an example.

Using the ACE framework to develop tests in multiple domains could provide additional evidence for the validity of the framework and, more importantly, could provide practical instruments allowing educators to assess important aspects of the scientific competencies of their students.

We conclude with directions for future research. First, the domain-general ACE subtest needs further validation. In this study we performed comparative studies with the domain-specific ACE subtest. Overall, the performance level of the learners in all modes of instruction was moderate on the domain-specific subtest, indicating a moderate level of modeling proficiency. It can be expected that modeling proficiency may grow with greater experience. At that point, as modeling proficiency grows, the domain-specific reasoning skills may be generalized to other domains and it might become relevant to assess domain-general reasoning skills in a posttest. Future research should validate the dimension of domain-specificity.

Second, future research may further investigate the difference between the domain-specific and domain-general subtest. With item construction formats that can be applied to both domain-specific and domain-general situations, the characteristics of parallel items might be compared. An interesting research question is how domain-general and domain-specific skills are related and how the domain-general skills develop after domain-specific instruction.

A third suggestion for future research can be found in the development of instruction. Modeling is a method of inquiring about phenomena that invokes the acquisition of scientific reasoning skills. The different aspects of scientific reasoning skills are trained by modeling instruction. More specifically, the aspects of scientific reasoning can be acquired through different modes of instruction (e.g., simulation-based inquiry and modeling). However, some aspects of modeling are not easily learned. The revised instruction in study 3 resulted in an improvement in reasoning with the relations in a model, but the create skills did not improve. Therefore, future research should investigate methods to acquire a higher level of modeling proficiency. The findings of the comparative studies showed that students did not easily acquire modeling skills at the highest level and that the modelers had relatively low create abilities. A limiting factor might have been the restricted amount of time that was available in the experiments. In order to achieve a higher level of modeling proficiency, it may be necessary to take the development of modeling instruction to the curriculum level, developing dedicated programs to acquire modeling competency, allowing students to spend more time on creating models and on applying models in various contexts.

In sum, the research presented in this dissertation took a promising step toward the clarification of the learning goals of modeling and the assessment of modeling skills. By classifying and measuring the learning outcomes of modeling, the ACE framework and assessment instrument contribute to an evidence-based discussion about the effectiveness of modeling-based inquiry.

Summary

Introduction

The topic of this thesis is computer modeling. A defining characteristic of computer modeling is the construction and exploration of executable computer models of dynamic systems. Modeling is a topic that aligns well with the current trend in secondary science education to actively involve students in their knowledge construction, give room for inquiry, and offer realistic tasks. It connects to the current reform in science education that stresses new learning outcomes such as scientific literacy, inquiry learning, and a hands-on and minds-on mentality (van Driel et al., 2001).

The effectiveness of new modes of instruction that aim at new learning outcomes has been debated by many researchers and educators. An example of a means for achieving new learning outcomes is *inquiry learning*. There are researchers who state that this form of learning is ineffective and that direct instruction would be a more effective approach (e.g., Kirschner et al., 2006; Mayer, 2004). It must be noted that this criticism concerns mainly unguided inquiry learning. There is evidence that learning by inquiry, especially in a guided way, provides a motivating and engaging form of learning (e.g., Kuhn et al., 2000).

The debate between educational reformers and their critics all too often remains undecided, due to a lack of rigorous evidence. Claims by either side are based mainly on studies that are hard to compare objectively due to a lack of standardized assessment. There is a need for a clear definition of the desired learning outcomes and suitable assessment tools. With such definitions and valid measurement instruments it is possible to assess and compare the effectiveness of different modes of instruction.

There are several claims with respect to the learning outcomes of computer modeling. It has been claimed that modeling contributes to the acquisition of *scientific reasoning skills* (Doerr, 1997; Van Joolingen & De Jong, 1997; Löhner et al., 2005; Stratford et al., 1998; Wells et al., 1995), *conceptual knowledge* (Jonassen et al., 2005; Nersessian, 1999; Stratford et al., 1998), and improves *insight into dynamic systems* (Booth Sweeney & Sterman, 2000; Hagmayer & Waldmann, 2000; Hmelo-Silver et al., 2007; Sterman, 1994; Wilensky & Resnick, 1999).

In this thesis the learning outcomes of computer modeling are defined and an assessment instrument is developed and validated.

Research question

The main research question in this dissertation was:

What specific learning outcomes can be expected from computer modeling and how can they be measured?

To answer this question, the reasoning processes involved in computer modeling were investigated. The findings of this exploration were systematized as what is called here the 'ACE framework' that describes modeling knowledge in three dimensions: 'type of reasoning', 'complexity', and 'domain-specificity'. The dimension of 'types of reasoning' includes *applying* (A), *creating* (C), and *evaluating* (E). The dimension of complexity distinguishes between reasoning with simple and complex situations. The dimension of domain-specificity describes the extent to which reasoning is dependent on the domain and distinguishes between domain-specific and domain-general.

The ACE test was developed based on the ACE framework with test items for each of the types of reasoning as applied in simple and complex situations. In addition, the ACE test contains items that assess conceptual knowledge about the domain. The dimension of domain-specificity was implemented in the specific domain of global warming. The question of how to validly measure the learning outcomes was addressed in a validation study that investigated construct validity. Validity was further explored by analyzing the discriminative power of the test in two studies comparing computer modeling with another mode of instruction. Finally, the data of these comparative studies were analyzed in a combined analysis to compare the modes of instruction across the studies.

Studies

Validation study

In the validation study (see Chapter 3) the ACE test was validated by administering the test to 131 students with different levels of modeling proficiency (eleventh grade students, first-year psychology students, and first-year students of engineering physics who had completed a course about modeling). An analysis of the responses with item response models yielded evidence that the reasoning skills are composed of four unidimensional subscales, namely Applying, Creating, Evaluating, and Reproducing conceptual knowledge. The first three subscales are the reasoning processes as described in the ACE framework. The fourth subscale, Reproducing, concerns the test items about conceptual knowledge of the domain. With regard to the dimension of complexity it was found that the test items about complex situations were significantly more difficult than the items about simple situations. The test appeared to discriminate between students with and without modeling experience and a trend was found with respect to the domain-specific subtest in favor of students with domain knowledge.

In order to further investigate the discriminative power of the test, two comparative studies were performed in which a modeling-based mode of instruction was compared with another mode of instruction. The main question is whether the test is sensitive enough to detect differences in learning outcomes between the modes of instruction. First, we compared modeling to an expository mode of instruction offering quite a contrast. Second, modeling was compared with simulation-based instruction that had more in common with modeling.

The first comparative study

In the first comparative study (see Chapter 4) we compared modeling to expository instruction. In both conditions students worked independently on assignments about global warming during two sessions. Students in the expository mode of instruction were directly exposed to the domain concepts. Students in the modeling condition worked with the concepts by creating a model of the domain. The performance on the domain-specific test was analyzed with the domain-general pretest scores as a covariate.

The principal finding of this study was that the modeling students performed better on problems for complex situations. More specifically, the learners in the modeling condition performed better on reproducing complex conceptual knowledge and on evaluating complex models and data. No differences were found for applying knowledge and creating models. These results indicate that the ACE test is able to measure differences between groups with respect to processes described in the ACE framework. As a consequence of the lack of difference for applying and creating along with the low performance for these processes the modeling instruction was revised.

The second comparative study

In the second comparative study (see Chapter 5) we compared modeling to simulation-based instruction. The modeling instruction was revised to put more emphasis on reasoning with the basic building blocks of modeling. Similar to the previous comparative study, students worked independently on assignments about global warming. The students in the simulation-based mode of instruction answered questions using a given simulation in which they could change parameters and observe the consequences. The students in the modeling mode of instruction constructed and explored a model.

As in the first comparative study the modeling group performed better on complex problems. Test performance across the conditions in this study showed differences for the processes of reproducing, applying, and creating. The students in the simulation-based mode of instruction performed better on reproducing simple conceptual knowledge. The modeling group performed better on applying complex knowledge and creating simple models. No differences were found for evaluating models and data. This result was consistent with our expectations, because both a model and a simulation can be used for experimentation and can support the acquisition of evaluating skills.

Combined analysis

The two comparative studies each compared two modes of instruction and provided detailed information about the differences in performance between two groups. The data of the two comparative studies were analyzed in a combined analysis (see Chapter 6) to further explore the advantages and disadvantages of the different modes of instruction. The question that is addressed is how the four conditions performed in relation to each other, such as the simulation-based condition compared with the expository condition. We are particularly interested in the difference between the two modeling conditions; in other words, whether the revised modeling instruction had been effective. The purpose of the revised instruction was to offer the modeling students more exercise in reasoning with the building blocks of a model. Our assumption was that the exercises would result in better reasoning skills and a higher level of modeling proficiency.

The studies were comparable because they had been conducted in comparable circumstances. The students were from the same two schools and the same three teachers. In the second comparative study, a fourth teacher's class participated in the experiment (with 24 out of 76 students). The comparability was confirmed by the fact that there were no significant differences for the pretest scores between the conditions in both studies.

The combined analysis revealed differences between the conditions with regard to several types of reasoning. First, students in both the expository and simulation-based modes of instruction performed significantly better on reproducing simple conceptual knowledge than those in the modeling condition in the second comparative study. This finding suggests that simple conceptual knowledge is best acquired without the task of creating a model. Second, the students in the revised modeling condition performed better on applying complex knowledge than all other conditions. This finding suggests that the revised instruction successfully focused on profound and well-reasoned modeling behavior resulting in better applying reasoning skills. Third, students in the expository mode of instruction performed significantly worse on the complex evaluate subscale than those in the modeling condition with the revised instruction. The performance of the other conditions suggests a trend that complex evaluate skills are acquired equally well in both a modeling-based and a simulation-based mode of instruction. Finally, no significant differences were found for the subscale of creating models.

A more detailed analysis of the answer elements with respect to the use of domain knowledge and the occurrence of errors in creating revealed a possible relation between references to domain knowledge and the acquired level of modeling proficiency. In general, domain knowledge is useful in effective problem solving, but in the current task setting using domain knowledge might also possibly be a strategy that hinders the application of reasoning skills. The analysis of these elements indicates that the students in the simulation-based mode of instruction referred significantly more to domain knowledge and made more errors in creating than the other conditions. This result might suggest that there is a higher tendency to use domain knowledge in cases of lower reasoning abilities. In other words, students lacking model-based reasoning skills more often try to fall back on their existing domain knowledge.

With respect to the effect of the revised instruction a salient difference was found for the complex apply items: the modelers who received the revised instruction performed significantly better than the modelers in the first comparative study and than students in the other conditions.

General conclusion

The main question in the research was: What specific learning outcomes can be expected from computer modeling and how can these be measured? The results of the three studies show that the core reasoning processes of modeling are validly described by the processes of applying, creating, evaluating, and reproducing. A general pattern that arose from the two comparative studies is the difference in performance between the groups with respect to simple and complex items and processes. Overall, the students in the modeling-based mode of instruction performed better on the complex items. The modeling processes are progressively more demanding with regard to the types of reasoning involved, from the straightforward *reproducing* of conceptual knowledge to the higher-order processes of *applying*, *creating*, and *evaluating*. The findings indicate a benefit for the learners in the expository and simulation-based instruction on the simplest process of reproducing conceptual knowledge. The modelers in the second comparative study benefited progressively more on the complex items and the more demanding processes of applying and creating. However, neither group reached a high level for the process of evaluating.

The following recommendations can be made for the most effective mode of instruction with respect to a particular learning effect. In order to learn simple conceptual domain knowledge, creating a model seems to be disadvantageous. The best performance appears to be realized in an expository or a simulation-based mode of instruction. However, in order to learn reasoning with complex structures, modeling seems to be most effective. Evaluating skills appear to be equally well acquired in both simulation-based and modeling-based modes of instruction. An expository mode of instruction appears to be less effective for these. Finally, with respect to create abilities no differences were found between the different modes of instruction. The lack of difference between the conditions with respect to create skills calls for future research. The revised instruction in the second comparative study resulted in an improvement in reasoning with complex relations in a model, but the create skills did not improve. Future research should investigate the design and evaluation of methods for acquiring a higher level of modeling proficiency.

Samenvatting

Dutch summary

Introductie

Het onderwerp van dit proefschrift is computermodelleren. Modelleren is hierbij gedefinieerd als het met behulp van de computer construeren van modellen van dynamische systemen. Het onderwerp modelleren sluit nauw aan bij de huidige trend in het middelbaar natuurwetenschappelijk onderwijs om leerlingen waar mogelijk zelf te laten ontdekken, actief te betrekken bij kennisconstructie en leeromgevingen aan te bieden die aansluiten bij de beroepspraktijk van wetenschappers. Het sluit tevens aan bij nieuwe leerdoelen zoals wetenschappelijke geletterdheid, onderzoekend leren en een actieve betrokkenheid.

De effectiviteit van de manier waarop deze nieuwe leerdoelen in de onderwijspraktijk gebracht worden, is onderwerp van discussie onder wetenschappers en in het publieke domein. Een voorbeeld van een actieve leermethode in het middelbaar onderwijs is onderzoekend leren. Er zijn stemmen die beweren dat dit per definitie een ineffectieve leermethode is (Kirschner et al., 2006). Daarbij moet wel opgemerkt worden dat deze kritiek vooral betrekking heeft op onbegeleid onderzoekend leren. Er is evidentie dat begeleid onderzoekend leren juist een effectieve leermethode is die meer leerwinst oplevert dan 'puur' onderzoekend leren (Mayer, 2004).

De claims die in deze discussie naar voren worden gebracht, zijn veelal gebaseerd op studies die moeilijk te vergelijken zijn door gebrek aan gestandaardiseerde toetsen. Het ontbreekt vaak aan duidelijkheid over de gewenste leeruitkomsten en aan meetmethoden om die te meten. Als de leeruitkomsten helder zijn gedefinieerd en meetbaar gemaakt door middel van valide meetinstrumenten, is het mogelijk om de effectiviteit van verschillende leermethoden te toetsen en te vergelijken.

Wat betreft modelleren zijn er leeruitkomsten geclaimd op verschillende gebieden. Er wordt geclaimd dat modelleren bijdraagt aan *vaardigheid in het wetenschappelijk redeneren* (Doerr, 1997; Van Joolingen & De Jong, 1997; Löhner et al., 2005; Stratford et al., 1998; Wells et al., 1995), aan *domeinkennis* (Jonassen et al., 2005; Nersessian, 1999; Stratford et al., 1998) en aan *inzicht in dynamische systemen* (Booth Sweeney & Sterman, 2000; Hagmayer & Waldmann, 2000; Hmelo-Silver et al., 2007; Sterman, 1994; Wilensky & Resnick, 1999).

In dit proefschrift zijn de leeruitkomsten van modelleren in kaart gebracht en is een toetsinstrument ontwikkeld en gevalideerd.

Onderzoeksvraag

De hoofdvraag in dit onderzoek was:

Welke specifieke leeruitkomsten kunnen verwacht worden van computermodelleren en hoe kunnen deze gemeten worden?

Om deze vraag te beantwoorden, werden de redeneerprocessen bestudeerd die optreden bij modelleren. De bevindingen werden samengevat in het zogenaamde ACE-raamwerk dat leeruitkomsten van modelleren beschrijft in drie dimensies: 'type redeneren', 'complexiteit' en 'domeinspecificiteit'. In het type redeneren onderscheiden we: toepassen ('Apply', A), creëren ('Create', C) en evalueren ('Evaluate', E). Binnen de dimensie complexiteit wordt redeneren met enkelvoudige en complexe situaties onderscheiden. De dimensie domeinspecificiteit beschrijft de mate waarin de redeneervaardigheid gebonden is aan het domein en onderscheidt domeinspecifiek en domeinonafhankelijk.

Op basis van het ACE-raamwerk werd de ACE-toets ontwikkeld met toetsvragen voor iedere combinatie van dimensies. Bovendien bevat de ACE-toets vragen naar conceptuele kennis over het domein. De dimensie van domeinspecificiteit was geïmplementeerd voor het specifieke domein van de opwarming van de aarde. De vraag hoe de leeruitkomsten valide gemeten kunnen worden, werd beantwoord in een validatiestudie waarin constructvaliditeit werd onderzocht. Vervolgens werd de validiteit verder onderzocht door het discriminerend vermogen van de toets te analyseren in twee studies waarin telkens twee manieren van instructie werden vergeleken. De gegevens van de twee vergelijkende studies werden tot slot naast elkaar gelegd om de condities tussen de twee studies te kunnen vergelijken.

Studies

Validatiestudie

In de validatiestudie (zie hoofdstuk 3) werd de ACE-toets gevalideerd door hem af te nemen bij 131 leerlingen en studenten die een verschillend niveau van modelleerervaring hadden (5 VWO scholieren, eerstejaars psychologiestudenten en eerstejaars natuurkundestudenten die een cursus modelleren hadden gevolgd). Een analyse van de antwoorden met item-responsemodellen toonde aan dat de redeneervaardigheden, kunnen worden onderverdeeld in vier eendimensionale subschalen, namelijk Toepassen, Creëren, Evalueren en Reproduceren van conceptuele kennis. De eerste drie subschalen zijn de redeneerprocessen beschreven in het ACE-raamwerk. De vierde subschaal, Reproduceren, heeft betrekking op de toetsvragen over conceptuele kennis van het domein. Met betrekking tot de dimensie complexiteit bleek dat de toetsvragen over de complexe situaties significant moeilijker waren dan de vragen over enkelvoudige situaties. De toets bleek te onderscheiden tussen de leerlingen met en zonder modelleerervaring en voor de scores op de domeinspecifieke toets was een trend te zien ten gunste van leerlingen met domeinkennis.

Om het onderscheidend vermogen van de toets verder te onderzoeken, zijn twee vergelijkende studies uitgevoerd waarin een op modelleren gebaseerde instructie is vergeleken met andere vormen van instructie. De hoofdvraag is of de toets in staat is verschillen in leeropbrengst tussen de instructiemethoden te detecteren. We vergeleken modelleren eerst met een instructievorm die sterk verschilt, namelijk directe instructie. Daarna vergeleken we modelleren met de verwante simulatiegebaseerde instructie.

De eerste vergelijkende studie

In de eerste vergelijkende studie (zie hoofdstuk 4) vergeleken we een modelleerinstructie met directe instructie. De leerlingen in beide condities werkten in twee sessies zelfstandig aan opdrachten over het klimaat. De leerlingen die directe instructie ontvingen, kregen de concepten op een directe manier aangeboden. De leerlingen in de modelleerconditie werkten met de concepten door middel van het maken van een model. De prestaties op de domeinspecifieke toets werden vergeleken waarbij de voortoetsscores voor de domeinonafhankelijke toets als covariaat werden gebruikt.

De belangrijkste bevinding van deze studie was dat de leerlingen in de modelleergroep beter presteerden op de complexe problemen. Preciezer, de modelleerders presteerden beter op het reproduceren van complexe conceptuele kennis en op het evalueren van complexe modellen en gegevens. Er werden geen verschillen gevonden voor het toepassen van kennis en het creëren van modellen. Deze resultaten tonen aan dat de ACE-toets in staat is om verschillen tussen groepen te meten met betrekking tot een aantal processen beschreven in het ACE-raamwerk. Het feit dat de processen toepassen en creëren geen verschillen lieten zien en dat beide groepen laag presteerden, was aanleiding om de instructie te reviseren.

De tweede vergelijkende studie

In de tweede vergelijkende studie (zie hoofdstuk 5) vergeleken we een modelleerinstructie met simulatiegebaseerd leren. De modelleerinstructie in deze studie legde meer nadruk op het redeneren met een model en op de basisbouwstenen van een model. Net als in de vorige vergelijkende studie werkten de leerlingen zelfstandig aan opdrachten over het klimaat. De leerlingen in de simulatiegebaseerde instructie maakten de opdrachten gebruikmakend van een gegeven simulatie waarvan ze parameters konden veranderen en de gevolgen daarvan konden observeren. De leerlingen in de op modelleren gebaseerde instructie construeerden zelf een model.

Net als in de eerste vergelijkende studie presteerde de modelleergroep beter op de complexe problemen. De toetsresultaten van beide condities laten ditmaal verschillen zien voor de processen reproduceren, toepassen en creëren. De leerlingen in de simulatiegroep presteerden beter op het reproduceren van enkelvoudige conceptuele kennis. De modelleergroep presteerde beter op het toepassen van complexe kennis en het creëren van simpele modellen. Voor de toetsvragen over het evalueren van modellen en experimentele data werden geen verschillen gevonden. Dit resultaat was in overeenstemming met onze verwachtingen, omdat zowel met een model als met een simulatie kan worden geëxperimenteerd en evalueervaardigheden kunnen worden verworven.

Gecombineerde analyse

De twee afzonderlijke studies vergeleken telkens twee condities en gaven gedetailleerde informatie over de prestatieverschillen tussen de groepen. Om een vollediger beeld te krijgen van de voor- en nadelen van de verschillende manieren van instructie voegden we de gegevens van de twee vergelijkende studies samen en voeren een gecombineerde analyse uit (zie hoofdstuk 6). De vraag die hiermee beantwoord kan worden, is hoe alle vier condities gepresteerd hebben ten opzichte van elkaar, bijvoorbeeld de simulatieconditie ten opzichte van de leerlingen die directe instructie ontvingen. In het bijzonder zijn we geïnteresseerd in het verschil tussen de twee modelleercondities, met andere woorden of de aanpassingen in de modelleerinstructie effect hebben gehad. De aangepaste instructie had vooral ten doel dat de modelleerders meer oefening zouden krijgen in het redeneren met de bouwstenen van een model. De aanname was dat deze oefening zou resulteren in betere redeneervaardigheden en een hoger algemeen modelleerniveau.

De studies konden vergeleken worden omdat de twee studies in overeenkomstige omstandigheden zijn uitgevoerd. De leerlingen kwamen van dezelfde twee scholen. Klassen van dezelfde drie docenten namen deel aan het onderzoek. In de tweede vergelijkende studie was nog een vierde docent betrokken bij het onderzoek (met 24 van de 76 leerlingen). De vergelijkbaarheid werd bevestigd door het feit dat de voortoetsscores geen significante verschillen lieten zien tussen de leerlingen van beide studies.

Verschillen tussen de condities werden gevonden met betrekking tot diverse typen redeneren. Ten eerste, voor het reproduceren van simpele conceptuele kennis bleken zowel de leerlingen die directe instructie als de leerlingen die simulatiegebaseerde instructie ontvingen significant beter te presteren dan de modelleerconditie in de tweede vergelijkende studie. Deze bevinding suggereert dat simpele conceptuele kennis het best geleerd wordt in een leeromgeving zonder de taak om een model te construeren. Ten tweede, wat betreft het toepassen van complexe kennis presteerden de leerlingen in de gereviseerde modelleerinstructie beter dan alle andere condities. Dit resultaat suggereert dat de gereviseerde instructie op een succesvolle manier meer nadruk legt op weldoordacht en beredeneerd modelleergedrag en leidt tot betere redeneervaardigheden op het gebied van het toepassen van kennis. Ten derde, voor de redeneervaardigheden met betrekking tot complex evalueren bleken de leerlingen die directe instructie ontvingen significant minder te presteren dan de modelleergroep in de gereviseerde instructie. De prestaties van de andere condities suggereren de trend dat evalueervaardigheden het best verworven kunnen worden in zowel een op modelleren gebaseerde als een simulatiegebaseerde instructie. Tot slot, voor de subschaal van het creëren van modellen werden onder de vier condities geen significante verschillen gevonden.

Naast het vergelijken van de prestaties op de deeltoetsen is ook een frequentie-analyse uitgevoerd van elementen in de toetsantwoorden die een indicatie zouden kunnen geven van het algemene modelleerniveau, namelijk het refereren aan domeinkennis en het maken van creëerfouten. Domeinkennis is in het algemeen weliswaar noodzakelijk voor een effectieve probleemaanpak, maar in de huidige taaksetting zou het gebruik van domeinkennis er ook toe kunnen leiden dat leerlingen minder gebruik maken van hun redeneervaardigheden. De analyse van deze elementen laat de trend zien dat de leerlingen in de simulatiegebaseerde conditie significant meer aan domeinkennis refereerden in hun antwoorden en meer creëerfouten maakten dan andere condities. Dit resultaat zou erop kunnen duiden dat leerlingen die in staat zijn om redeneervaardigheden toe te passen minder neiging hebben om domeinkennis te gebruiken en dat, omgekeerd, leerlingen met minder redeneervaardigheden vaker terugvallen op domeinkennis.

Met betrekking tot het effect van de revisie van de modelleerinstructie werd een opvallend verschil gevonden voor het toepassen van complexe kennis: de modelleergroep in de gereviseerde instructie haalde significant betere toetsresultaten dan de modelleergroep uit de eerste vergelijkende studie en ook dan de overige groepen.

Algemene conclusie

De hoofdvraag van dit onderzoek was: Welke specifieke leeruitkomsten kunnen verwacht worden van computer modelleren en hoe kunnen deze gemeten worden? De resultaten van de drie studies laten zien dat de kernprocessen van computermodelleren valide beschreven worden door de processen toepassen, creëren, evalueren en reproduceren. De vergelijking van verschillende manieren van instructie laat een patroon zien van verschillen wat betreft simpele en complexe toetsvragen en redeneerprocessen. De modelleergroep presteerde in beide vergelijkende studies beter op de complexe toetsvragen. Ook bij de redeneerprocessen is de trend zichtbaar dat modelleren vooral leerwinst oplevert voor de complexere processen. De redeneerprocessen hebben een oplopende graad van complexiteit, van het eenvoudige *reproduceren* van conceptuele kennis tot de steeds complexere processen *toepassen*, *creëren* en *evalueren*. De leerlingen in de directe en simulatiegebaseerde instructie presteerden het best op toetsvragen over het reproduceren van simpele conceptuele kennis. De modelleerders daarentegen hadden meer profijt voor de complexe toetsvragen en voor de complexere processen toepassen en creëren. Geen van beide groepen behaalde een hoog niveau voor het proces van evalueren.

De volgende aanbevelingen kunnen gedaan worden voor de meest effectieve manier van instructie met betrekking tot specifieke leerwinsten. Voor het leren van simpele conceptuele kennis lijkt het maken van een model nadelig. De hoogste leerwinst lijkt te worden gerealiseerd met directe of simulatiegebaseerde instructie. Echter, voor het leren toepassen van complexe modelstructuren lijkt modelleren de meest geschikte manier van instructie. Evalueervaardigheden worden in gelijke mate geleerd in zowel een op modelleren gebaseerde als een simulatiegebaseerde instructie. Directe instructie lijkt voor deze vaardigheden minder effectief. Tenslotte, met betrekking tot creëervaardigheden werden geen verschillen gevonden voor de verschillende manieren van instructie. Het gebrek aan verschillen tussen condities voor creëervaardigheden nodigt uit tot vervolgonderzoek. De gereviseerde instructie in de tweede vergelijkende studie resulteerde in een hoger redeneerniveau voor het toepassen van complexe modelstructuren, maar niet in betere creëervaardigheden. Het ontwerpen en evalueren van methoden waarmee een hoger modelleerniveau kan worden verworven, is een interessant onderwerp voor vervolgstudies.

References

Alessi, S. (2005). The application of system dynamics modeling in elementary and secondary school curricula. from http://www.c5.cl/ieinvestiga/actas/ribie2000/charlas/alessi.htm

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching and assessing: A revision of bloom's taxonomy of educational objectives*. New York: Longman.

Archbald, D. A., & Newmann, F. M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in the secondary school*. Reston, VA: National Association of Secondary School Principals.

Assaraf, O. B. Z., & Orion, N. (2005). Development of system thinking skills in the context of earth system education. *Journal of Research in Science Teaching, 42*, 518-560.

Baggott La Velle, L., McFarlane, A., & Brawn, R. (2003). Knowledge transformation through ICT in science education: A case study in teacher-driven curriculum development - case-study 1. *British Journal of Educational Technology, 34*, 183-199.

Blake, C., & Scanlon, E. (2007). Reconsidering simulations in science education at a distance: Features of effective use. *Journal of Computer Assisted Learning, 23*, 491-502.

Bliss, J., Ogborn, J., Boohan, R., Briggs, J., Brosnan, T., Brough, D., Mellar, H., Miller, R., Nash, C., Rodgers, C., & Sakonidis, B. (1992). Reasoning supported by computational tools. *Computers & Education, 18*, 1-9.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals, by a committee of college and university examiners. Handbook i: Cognitive domain*. New York: David McKay Company.

Booth Sweeney, L., & Sterman, J. D. (2000). Bathtub dynamics: Initial results of a systems thinking inventory. *System Dynamics Review, 16*, 249-286.

Booth Sweeney, L., & Sterman, J. D. (2007). Thinking about systems: Student and teacher conceptions of natural and social systems. *System Dynamics Review, 23*, 285-311.

Van Borkulo, S. P., Van Joolingen, W. R., Savelsbergh, E. R., & De Jong, T. (2008). A framework for the assessment of learning by modeling. In P. Blumschein, J. Stroebel, W. Hung & D. Jonassen (Eds.), *Model-based approaches to learning* (pp. 177-196). Rotterdam, Netherlands: Sense Publishers.

Bravo, C., Van Joolingen, W. R., & De Jong, T. (2006). Modeling and simulation in inquiry learning: Checking solutions and giving intelligent advice.

Simulation-Transactions of the Society for Modeling and Simulation International, 82, 769-784.

Buckley, B. C., Gobert, J. D., Kindfield, A. C. H., Horwitz, P., Tinker, R. F., Gerlits, B., Wilensky, U., Dede, C., & Willett, J. (2004). Model-based teaching and learning with biologica™: What do they learn? How do they learn? How do we know? *Journal of Science Education and Technology, 13*, 23-41.

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

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Chi, M. T. H., & Roscoe, R. D. (2002). The processes and challenges of conceptual change. In M. Limón & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice*.

Clement, J. (2000). Model based learning as a key research area for science education. *International Journal of Science Education, 22*, 1041-1053.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cronin, M. A., & Gonzalez, C. (2007). Understanding the building blocks of dynamic systems. *System Dynamics Review, 23*, 1-17.

Davies, P. (1999). What is evidence-based education? *British Journal of Educational Studies, 47*, 108-121.

Doerr, H. M. (1996). Stella ten years later: A review of the literature. *International journal of computers for mathematical learning, 1*, 201.

Doerr, H. M. (1997). Experiment, simulation and analysis: An integrated instructional approach to the concept of force. *International Journal of Science Education, 19*, 265-282.

van Driel, J. H., Beijaard, D., & Verloop, N. (2001). Professional development and reform in science education: The role of teachers' practical knowledge. *Journal of Research in Science Teaching, 38*, 137-158.

van Driel, J. H., & Verloop, V. (2002). Experienced teachers' knowledge of teaching and learning of models and modelling in science education. *International Journal of Science Education, 24*, 1255-1272.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Ergazaki, M., Komis, V., & Zogza, V. (2005). High-school students' reasoning while constructing plant growth models in a computer-supported educational environment. *International Journal of Science Education, 27*, 909-933.

Evans, J. S. B. T. (1989). *Bias in human reasoning: Causes and consequences*: Erlbaum.

Fiddick, L., Cosmides, L., & Tooby, J. (2000). No interpretation without representation: The role of domain-specific representations and inferences in the Wason selection task. *Cognition, 77*, 1-79.

Forbus, K. D. (1996). Why computer modeling should become a popular hobby. *D-Lib Magazine, 2*.

Forbus, K. D., Carney, K., Sherin, B. L., & Ureel, L. C. (2005). Vmodel - a visual qualitative modeling environment for middle-school students. *Ai Magazine, 26*, 63-72.

Forrester, J. W. (1994). *Learning through system dynamics as preparation for the 21st century*. Paper presented at the Systems Thinking and Dynamic Modeling Conference for K-12 Education. Concord, MA, USA. June 27-29, 1994.

Fretz, E. B., Wu, H. K., Zhang, B. H., Davis, E. A., Krajcik, J. S., & Soloway, E. (2002). An investigation of software scaffolds supporting modeling practices. *Research in Science Education, 32*, 567-589.

Geier, R., Blumenfeld, P. C., Marx, R. W., Krajcik, J. S., Fishman, B., Soloway, E., & Clay-Chambers, J. (2008). Standardized test outcomes for students engaged in inquiry-based science curricula in the context of urban reform. *Journal of Research in Science Teaching, 45*, 922-939.

Gentner, D., & Stevens, A. L. (Eds.). (1983). *Mental models*. Hillsdale, NJ: Erlbaum.

Glymour, C. N., & Cooper, G. F. (Eds.). (1999). *Computation, causation, and discovery*. Menlo Park, California: American Association for Artificial Intelligence Press.

Gomez, P. J. S. (2005). Defending constructivism in science education. *Revista Espanola De Pedagogia, 63*, 162-166.

Guttersrud, O. (2006). *Toward a description of physics students' modelling competency*. Paper presented at the GIREP. Amsterdam.

Hagmayer, Y., & Waldmann, M. R. (2000). *Simulating causal models: The way to structural sensitivity.* Paper presented at the Twenty-Second Annual Conference of the Cognitive Science Society.

Hanauer, D. I., Jacobs-Sera, D., Pedulla, M. L., Cresawn, S. G., Hendrix, R. W., & Hatfull, G. F. (2006). Teaching scientific inquiry. *Science, 314*, 1880-1881.

Harlen, W. (2001). The assessment of scientific literacy in the OECD/PISA project. In H. Behrendt, H. Dahncke, R. Duit, W. Gräber, M. Komorek,

A. Kross & P. Reiska (Eds.), *Research in science education - past, present, and future* (Vol. 36, pp. 49-60).

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*, 77-89.

Hennessy, S., Wishart, J., Whitelock, D., Deaney, R., Brawn, R., la Velle, L., McFarlane, A., Ruthven, K., & Winterbottom, M. (2007). Pedagogical approaches for technology-integrated science teaching. *Computers & Education, 48*, 137-152.

Hestenes, D. (1987). Toward a modeling theory of physics instruction. *American journal of physics, 55*, 440-454.

Hestenes, D. (1996). *Modeling methodology for physics teachers.* Paper presented at the Proceedings of the International Conference on Undergraduate Physics Education, College Park.

Hickey, D. T., Kindfield, A. C. H., Horwitz, P., & Christie, M. A. (1999). Advancing educational theory by enhancing practice in a technology-supported genetics learning environment. *Journal of Education, 181*, 25-55.

Hmelo-Silver, C. E., & Azevedo, R. (2006). Understanding complex systems: Some core challenges. *Journal of the Learning Sciences, 15*, 53-61.

Hmelo-Silver, C. E., Marathe, S., & Liu, L. (2007). Fish swim, rocks sit, and lungs breathe: Expert-novice understanding of complex systems. *Journal of the Learning Sciences, 16*, 307-331.

Hmelo-Silver, C. E., Nagarajan, A., & Day, R. S. (2002). "It's harder than we thought it would be": A comparative case study of expert-novice experimentation strategies. *Science Education, 86*, 219-243.

Hmelo, C. E., Holton, D. L., & Kolodner, J. L. (2000). Designing to learn about complex systems. *Journal of the Learning Sciences, 9*, 247-298.

Hogan, K., & Thomas, D. (2001). Cognitive comparisons of students' systems modeling in ecology. *Journal of Science Education and Technology, 10*, 319-345.

Holland, J. H. (2000). *Emergence: From chaos to order*: Oxford University Press.

Jackson, S. L., Stratford, S. J., Krajcik, J. S., & Soloway, E. (1996). Making dynamic modeling accessible to pre-college science students. *Interactive Learning Environments, 4*, 233-257.

Jacobson, M. J., & Wilensky, U. (2006). Complex systems in education: Scientific and educational importance and implications for the learning sciences. *Journal of the Learning Sciences, 15*, 11-34.

Johnson-Laird, P. N. (2001). Mental models and deduction. *Trends in Cognitive Sciences, 4*, 434-442.

Jonassen, D. H., Carr, C., & Yueh, H. P. (1998). Computers as mindtools for engaging learners in critical thinking. *TechTrends, 43*, 24-32.

Jonassen, D. H., Ströbel, J., & Gottdenker, J. (2005). Model building for conceptual change. *Interactive Learning Environments, 13*, 15-37.

De Jong, T. (2006). Computer simulations: Technological advances in inquiry learning. *Science, 312*, 532-533.

De Jong, T., & Van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research, 68*, 179-201.

De Jong, T., & Van Joolingen, W. R. (2007). Model-facilitated learning. In J. M. Spector, M. D. Merrill, J. J. G. van Merriënboer & M. P. Driscoll (Eds.), *Handbook of research on educational communication and technology* (3rd ed., pp. 457-468): Lawrence Erlbaum.

Van Joolingen, W. R., & De Jong, T. (1997). An extended dual search space model of scientific discovery learning. *Instructional Science, 25*, 307-346.

Van Joolingen, W. R., De Jong, T., & Dimitrakopoulou, A. (2007). Issues in computer supported inquiry learning in science. *Journal of Computer Assisted Learning, 23*, 111-119.

van Joolingen, W. R., de Jong, T., Lazonder, A. W., Savelsbergh, E. R., & Manlove, S. (2005). Co-lab: Research and development of an online learning environment for collaborative scientific discovery learning. *Computers in Human Behavior, 21*, 671-688.

Kainz, D., & Ossimitz, G. (2002). *Can students learn stock-flow-thinking? An empirical investigation.* Paper presented at the 2002 System Dynamics Conference. Palermo, Italy.

Kauertz, A., & Fischer, H. E. (2006). Assessing students' level of knowledge and analysing the reasons for learning difficulties in physics by Rasch analysis. In X. Liu & W. J. Boone (Eds.), *Applications of Rasch measurement in science education* (pp. 212–246). Maple Grove, Minnesota: JAM Press.

Kennedy, C., Wilson, M., Draney, K., Tutunciyan, S., & Vorp, R. (2008). Constructmap (version 4.3.14) [computer program].

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*, 75-86.

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

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

Komis, V., Ergazaki, M., & Zogza, V. (2007). Comparing computer-supported dynamic modeling and paper & pencil' concept mapping technique in students' collaborative activity. *Computers & Education, 49*, 991-1017.

Kruckeberg, R. (2006). A deweyan perspective on science education: Constructivism, experience, and why we learn science. *Science & Education, 15*, 1-30.

Kuhn, D. (2007). Reasoning about multiple variables: Control of variables is not the only challenge. *Science Education, 91*, 710-726.

Kuhn, D., Black, J., Keselman, A., & Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cognition and Instruction, 18*, 495-523.

Lederman, J., Lederman, N., Wickman, P.-O., & Lager-Nyqvist, L. (2007). *An international, systematic investigation of the relative effects of inquiry and direct instruction*. Paper presented at the ESERA. Malmö.

Linn, M. C., Lee, H. S., Tinker, R., Husic, F., & Chiu, J. L. (2006). Teaching and assessing knowledge integration in science. *Science, 313*, 1049-1050.

Liu, X., & Boone, W. (2006). Introduction to Rasch measurement in science education. In X. Liu & W. Boone (Eds.), *Applications of Rasch measurement in science education*. Maple Grove, Minnesota: JAM Press.

Löhner, S. (2005). *Computer based modeling tasks: The role of external representation.* University of Amsterdam, Amsterdam.

Löhner, S., Van Joolingen, W. R., & Savelsbergh, E. R. (2003). The effect of external representation on constructing computer models of complex phenomena. *Instructional Science, 31*, 395-418.

Löhner, S., van Joolingen, W. R., Savelsbergh, E. R., & van Hout-Wolters, B. (2005). Students' reasoning during modeling in an inquiry learning environment. *Computers in Human Behavior, 21*, 441-461.

Lopes, J. B., & Costa, N. (2007). The evaluation of modelling competences: Difficulties and potentials for the learning of the sciences. *International Journal of Science Education, 29*, 811-851.

Lynch, S., Kuipers, J., Pyke, C., & Szesze, M. (2005). Examining the effects of a highly rated science curriculum unit on diverse students: Results from a planning grant. *Journal of Research in Science Teaching, 42*, 912-946.

Maaß, K. (2006). What are modelling competencies? *ZDM, 38*, 113-142.

Magnani, L., Nersessian, N. J., & Thagard, P. (Eds.). (1998). *Model-based reasoning in scientific discovery*. New York: Kluwer Academic/Plenum Publishers.

Mandinach, E. B. (1989). Model-building and the use of computer-simulation of dynamic-systems. *Journal of Educational Computing Research, 5*, 221-243.

Mandinach, E. B., & Cline, H. F. (1996). Classroom dynamics: The impact of a technology-based curriculum innovation on teaching and learning. *Journal of Educational Computing Research, 14*, 83-102.

Manlove, S., Lazonder, A. W., & De Jong, T. (2006). Regulative support for collaborative scientific inquiry learning. *Journal of Computer Assisted Learning, 22*, 87-98.

Masood, E. (1998). Uk seeks physicists for environmental research. *Nature, 393*, 400.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *American Psychologist, 59*, 14-19.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (third ed., pp. 13-103). Washington, DC: American Council on Education and the National Council on Measurement in Education.

Milrad, M., Spector, J. M., & Davidson, P. I. (2003). Model facilitated learning. In S. Naidu (Ed.), *Learning & teaching with technology: Principles and practices* (pp. 13-28). London: Kogan Page.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Nersessian, N. J. (1999). Model-based reasoning in conceptual change. In L. Magnani, N. J. Nersessian & P. Thagard (Eds.), *Model-based reasoning in scientific discovery* (pp. 5-22). New York: Kluwer Academic/Plenum Publishers.

Novak, J. D. (1990). Concept mapping - a useful tool for science-education. *Journal of Research in Science Teaching, 27*, 937-949.

Papaevripidou, M. M., Constantinou, C. C. P., & Zacharia, Z. Z. C. (2007). Modeling complex marine ecosystems: An investigation of two teaching approaches with fifth graders. *Journal of Computer Assisted Learning, 23*, 145.

Papert, S., & Harel, I. (1991). Situating constructionism. In S. Papert & I. Harel (Eds.), *Constructionism* (pp. 1-11): Ablex Publishing Corporation.

Park, H. R. (2008). ICT in science education: A quasi-experimental study of achievement, attitudes toward science, and career aspirations of korean middle school students. *International Journal of Science Education*, 1-20.

Polit, D. F., & Hungler, B. P. (1991). *Nursing research: Principles and methods*. Philadelphia, PA: Lippincott.

Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., Kyza, E., Edelson, D., & Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *Journal of the Learning Sciences, 13*, 337-386.

Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology, 99*, 561-574.

Royer, J. M., Cisero, C. A., & Carlo, M. S. (1993). Techniques and procedures for assessing cognitive skills. *Review of Educational Research, 63*, 201-243.

Sabelli, N. H. (2006). Complexity, technology, science, and education. *Journal of the Learning Sciences, 15*, 5-9.

Sato, T. (1996). Type I and type II error in multiple comparisons. *Journal of Psychology, 130*, 293-302.

Schecker, H. P. (1998). *Physik – modellieren, grafikorientierte modelbildungssysteme im physikunterricht*. Stuttgart, Germany: Ernst Klett Verlag GmbH.

Schecker, H. P., & Einhaus, E. A. (2007). *Modelling science competencies*. Paper presented at the ESERA. Malmö. August 2007.

Schwartz, R. S., Lederman, N. G., & Crawford, B. A. (2004). Developing views of nature of science in an authentic context: An explicit approach to bridging the gap between nature of science and scientific inquiry. *Science Education, 88*, 610-645.

Schwarz, C. V., & White, B. Y. (2005). Metamodeling knowledge: Developing students' understanding of scientific modeling. *Cognition and Instruction, 23*, 165-205.

Sins, P. H. M. (2006). *Students' reasoning during computer-based scientific modeling.* University of Amsterdam.

Sins, P. H. M., Savelsbergh, E. R., & Van Joolingen, W. R. (2005). The difficult process of scientific modelling: An analysis of novices' reasoning during computer-based modelling. *International Journal of Science Education, 27*, 1695-1721.

Skemp, R. R. (2006). Relational understanding and instrumental understanding. *Mathematics Teaching in the Middle School, 12*.

Spector, J. M. (2000). System dynamics and interactive learning environments: Lessons learned and implications for the future. *Simulation & Gaming, 31*, 509-516.

Spector, J. M. (2001). Tools and principles for the design of collaborative learning environments for complex domains. *Journal of Structural Learning & Intelligent Systems, 14*, 483-510.

Spector, J. M., Christensen, D. L., Sioutine, A. V., & McCormack, D. (2001). Models and simulations for learning in complex domains: Using causal loop diagrams for assessment and evaluation. *Computers in Human Behavior, 17*, 517-545.

Star, J. R., & Rittle-Johnson, B. (2008). Flexibility in problem solving: The case of equation solving. *Learning and Instruction, 18*, 565-579.

Steed, M. (1992). Stella, a simulation construction kit: Cognitive process and educational implications. *Journal of Computers in Mathematics and Science Teaching, 11*, 39-52.

Sterman, J. D. (1994). Learning in and about complex systems. *System Dynamics Review, 10*, 291-330.

Sterman, J. D. (2002). All models are wrong: Reflections on becoming a systems scientist. *System Dynamics Review, 18*, 501-531.

Stouthard, M. E. A. (1998). Validiteit. In W. P. van den Brink & G. J. Mellenbergh (Eds.), *Testleer en testconstructie* (pp. 269-301). Amsterdam: Boom.

Stratford, S. J. (1997). A review of computer-based model research in precollege science classrooms. *Journal of Computers in Mathematics and Science Teaching, 16*, 3-23.

Stratford, S. J., Krajcik, J., & Soloway, E. (1998). Secondary students' dynamic modeling processes: Analyzing, reasoning about, synthesizing, and testing models of stream ecosystems. *Journal of Science Education and Technology, 7*, 215-234.

Swaak, J., & de Jong, T. (1996). Measuring intuitive knowledge in science: The development of the what-if test. *Studies In Educational Evaluation, 22*, 341-362.

Tobin, K., & Tippins, D. (1993). Constructivism as a referent for teaching and learning. In *The practice of constructivism in science education* (pp. 3-21).

Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 215-237). New York, NJ: Springer.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). One parameter logistic model OPLM (computer software maual). Arnhem, The Netherlands: CITO.

Wason, P. C. (1968). Reasoning about a rule. *QJ Exp Psychol, 20*, 273-281.

Wells, M., Hestenes, D., & Swackhamer, G. (1995). A modeling method for high-school physics instruction. *American Journal of Physics, 63*, 606-619.

Westbrook, L. (2006). Mental models: A theoretical overview and preliminary study. *Journal of Information Science, 32*, 563-579.

Wilensky, U., & Resnick, M. (1999). Thinking in levels: A dynamic systems approach to making sense of the world. *Journal of Science Education and Technology, 8*, 3-19.

Appendix

The standardized correct answer elements in the domain of the energy of the earth

Definitional

- albedo is the *part* of the radiation that is reflected
- depends on the color of the surface
- heat capacity is the amount of heat required to raise the temperature
- ... of a body
- ... by one degree Celsius
- reference to the definition of heat capacity
- reference to the definition of albedo
- reference to formula temperature = energy / heat capacity
- a constant temperature means inflow = outflow
- a constant temperature means equilibrium
- elaborate explanation of albedo (the higher albedo, the more reflection; black isn’t reflecting anything, albedo = 0%, white is reflecting all radiation, albedo = 100%)
- elaborate description of equilibrium

Relational

- existence of the relation between variable 1 and variable 2
- direction of the relation from variable 1 to variable 2
- quality of the relation between variable 1 and variable 2
- shortcut reasoning: expressing a correct indirect relation but leaving out steps in between
- apply a formula in a calculation (15000 J) /(50 J/K)
- temperature is increasing
- the increase is fast in the beginning and slow in the end
- explicit statement that temperature is reaching equilibrium
- equilibrium is reached within 3 years
- reference to the term ‘negative feedback’
- explaining negative feedback
- less energy is reaching the earth
- energy in the battery generates heat
- albedo will decrease
- there are too many factors involved to say something reasonable

Evaluative

- the conclusion is (not) supported by the data in the graph or table

(posing concrete evidence)

- all other variables must be the same in both experiments
- two variables have been changed at the same time
- two experiments is too little to draw any conclusion
- one variable has been changed
- reference to the data in the graph or table
- in the long term the relation is less strong
- temperature in both experiments reaches the same value
- temperature increases less fast when heat capacity is higher
- vary the settings of temperature
- vary the settings of heat capacity
- vary the settings of the radiation of the sun
- observe outgoing radiation, temperature or the results
- draw a conclusion
- the beginning of the graph line matches with the graph line of our model
- other variable(s) play a role
- ... from 2002
- description of a possible additional variable
- the oscillation need to be explained
- variable heat capacity is missing in the model
- variable energy is missing in the model
- outgoing radiation has an indirect relation with temperature
- inflow of energy has an indirect relation with temperature
- take the goal of the model into account
- there is no relation
- the relation is the other way round

Creational

- creation of an additional variable describing the new phenomenon
- creation of an additional relation between variable 1 and variable 2 (existence of the relation)
- creation of the direction of the relation from variable 1 to variable 2
- creation of the type of the relation between variable 1 and variable 2
- creation of an additional variable describing part of the new phenomenon
- creation of an extra additional variable for reasons of restructuring the model
- creation of an additional variable not necessary but functional

Curriculum vitae

Sylvia van Borkulo was born on April 27, 1972 in Amsterdam. In 1990 she finished her secondary education (gymnasium α) at the 'Waterlant-College' in Amsterdam and started to study mathematics at the VU University Amsterdam. In 1994 she combined the study mathematics with the study classical piano at the Utrecht Conservatoire. In 1996 she graduated in mathematics on a master thesis on decoding convolutional codes. In 1999 she graduated from the Utrecht Conservatoire with specializations Music Education and Ensemble Leading.

Sylvia worked as a software engineer at Cap Gemini from November 1999 until September 2001 and at the Academic Medical Center Amsterdam from January 2002 until May 2004. In parallel, she worked as a piano teacher for 7 years. In November 2003 she started to study educational sciences at the Utrecht University and obtained a bachelor's degree in November 2004.

From December 2004 Sylvia worked as a PhD student at the University of Twente, Faculty Behavioral Sciences, Department of Instructional Technology on the topic 'Assessment of learning outcomes of computer modeling environments'.

ICO dissertation series

In the ICO Dissertation Series dissertations are published of graduate students from faculties and institutes on educational research within the following universities: Eindhoven University of Technology, Leiden University, Maastricht University, Open University of the Netherlands, University of Amsterdam, University of Groningen, University of Twente, Utrecht University, VU University Amsterdam, and Wageningen University (and formerly Radboud University Nijmegen and Tilburg University).

150. Sins, P.H.M. (18-05-2006). *Students' reasoning during computer-based scientific modeling.* Amsterdam: University of Amsterdam.
151. Mathijsen, I.C.H. (24-05-2006). *Denken en handelen van docenten.* Utrecht: Utrecht University.
152. Akkerman, S.F. (23-06-2006). *Strangers in dialogue: Academic collaboration across organizational boundaries.* Utrecht: Utrecht University.
153. Willemse, T.M. (21-08-2006). *Waardenvol opleiden: Een onderzoek naar de voorbereiding van aanstaande leraren op hun pedagogische opdracht.* Amsterdam: VU University Amsterdam.
154. Kieft, M. (19-09-2006). *The effects of adapting writing instruction to students' writing strategies.* Amsterdam: University of Amsterdam.
155. Vreman-de Olde, G.C. (27-09-2006). *Look experiment design: Learning by designing instruction.* Enschede: University of Twente.
156. Van Amelsvoort, M. (13-10-2006). *A space for debate: How diagrams support collaborative argumentation-based learning.* Utrecht: Utrecht University.
157. Oolbekking-Marchand, H. (9-11-2006). *Teachers' perspectives on self-regulated learning: An exploratory study in secondary and university education.* Leiden: Leiden University.
158. Gulikers, J. (10-11-2006). *Authenticity is in the eye of the beholder: Beliefs and perceptions of authentic assessment and the influence on student learning.* Heerlen: Open University of the Netherlands.
159. Henze, I. (21-11-2006). *Science teachers' knowledge development in the context of educational innovation.* Leiden: Leiden University.
160. Van den Bossche, P. (29-11-2006). *Minds in teams: The influence of social and cognitive factors on team learning.* Maastricht: Maastricht University.
161. Mansvelder-Longayroux, D.D. (06-12-2006). *The learning portfolio as a tool for stimulating reflection by student teachers.* Leiden: Leiden University.
162. Visschers-Pleijers, A.J.S.F. (19-01-2007). *Tutorial group discussion in problem-based learning: Studies on the measurement and nature of learning-oriented student interactions.* Maastricht: Maastricht University.

163. Poortman, C.L. (16-02-2007). *Workplace learning processes in senior secondary vocational education.* Enschede: University of Twente.
164. Schildkamp. K.A. (15-03-2007). *The utilisation of a self-evaluation instrument for primary education.* Enschede: University of Twente.
165. Karbasioun, M. (20-04-2007). *Towards a competency profile for the role of instruction of agricultural extension professionals in Asfahan.* Wageningen: Wageningen University.
166. Van der Sande, R.A.W. (04-06-2007). *Competentiegerichtheid en scheikunde leren: Over metacognitieve opvattingen, leerresultaten en leeractiviteiten.* Eindhoven: Eindhoven University of Technology.
167. Pijls, M. (13-06-2007). *Collaborative mathematical investigations with the computer: Learning materials and teacher help.* Amsterdam: University of Amsterdam.
168. Könings, K. (15-06-2007). *Student perspectives on education: Implications for instructional design.* Heerlen: Open University of the Netherlands.
169. Prangsma, M.E. (20-06-2007). *Multimodal representations in collaborative history learning.* Utrecht: Utrecht University.
170. Niemantsverdriet, S. (26-06-2007). *Learning from international internships: A reconstruction in the medical domain.* Maastricht: Maastricht University.
171. Van der Pol, J. (03-07-2007). *Facilitating online learning conversations: Exploring tool affordances in higher education.* Utrecht: Utrecht University.
172. Korobko, O.B. (07-09-2007). *Comparison of examination grades using item response theory: A case study*. Enschede: University of Twente.
173. Madih-Zadeh, H. (14-09-2007). *Knowledge construction and participation in an asynchronous computer-supported collaborative learning environment in higher education.* Wageningen: Wageningen University.
174. Budé, L.M. (05-10-2007). *On the improvement of students' conceptual understanding in statistics education.* Maastricht: Maastricht University.
175. Meirink, J.A. (15-11-2007). *Individual teacher learning in a context of collaboration in teams.* Leiden: Leiden University.
176. Niessen, T.J.H. (30-11-2007). *Emerging epistemologies: Making sense of teaching practices.* Maastricht: Maastricht University.
177. Wouters, P. (07-12-2007). *How to optimize cognitive load for learning from animated models.* Heerlen: Open University of the Netherlands.
178. Hoekstra, A. (19-12-2007). *Experienced teachers' informal learning in the workplace.* Utrecht: Utrecht University.
179. Munneke-de Vries, E.L. (11-01-2008). *Arguing to learn: Supporting interactive argumentation through computer-supported collaborative learning.* Utrecht: Utrecht University.

180. Nijveldt, M.J. (16-01-2008). *Validity in teacher assessment. An exploration of the judgement processes of assessors*. Leiden: Leiden University.
181. Jonker, H.G. (14-02-2008). *Concrete elaboration during knowledge acquisition*. Amsterdam: VU University Amsterdam.
182. Schuitema, J.A. (14-02-2008). *Talking about values. A dialogue approach to citizenship education as an integral part of history classes*. Amsterdam: University of Amsterdam.
183. Janssen, J.J.H.M. (14-03-2008). *Using visualizations to support collaboration and coordination during computer-supported collaborative learning*. Utrecht: Utrecht University.
184. Honingh, M.E. (17-04-2008). *Beroepsonderwijs tussen publiek en privaat: Een studie naar opvattingen en gedrag van docenten en middenmanagers in bekostigde en niet-bekostigde onderwijsinstellingen in het middelbaar beroepsonderwijs.* Amsterdam: University of Amsterdam.
185. Baartman, L.K.J. (24-04-2008). *Assessing the assessment: Development and use of quality criteria for competence assessment programmes*. Utrecht: Utrecht University.
186. Corbalan Perez, G. (25-04-2008). *Shared control over task selection: Helping students to select their own learning tasks*. Heerlen: Open University of the Netherlands.
187. Hendrikse, H.P. (22-05-2008). *Wiskundig actief: Het ondersteunen van onderzoekend leren in het wiskunde onderwijs*. Enschede: University of Twente.
188. Moonen, M.L.I. (26-09-2008). *Testing the multi-feature hypothesis: Tasks, mental actions and second language acquisition*. Utrecht: Utrecht University.
189. Hooreman, R.W. (18-11-2008). *Synchronous coaching of the trainee teacher: An experimental approach*. Eindhoven: Eindhoven University of Technology.
190. Bakker, M.E.J. (02-12-2008). *Design and evaluation of video portfolios: Reliability, generalizability, and validity of an authentic performance assessment for teachers.* Leiden: Leiden University.
191. Kicken, W. (12-12-2008). *Portfolio use in vocational education: Helping students to direct their learning*. Heerlen: Open University of the Netherlands.
192. Kollöffel, B.J. (18-12-2008). *Getting the picture: The role of external representations in simulation-based inquiry learning*. Enschede: University of Twente.
193. Walraven, A. (19-12-2008). *Becoming a critical websearcher: Effects of instruction to foster transfer.* Heerlen: Open University of the Netherlands.