Paper presented at the 29th International Conference of the System Dynamics Society

July 24-28, 2011, Washington, DC

Measuring Knowledge Acquisition in Dynamic Decision Making Tasks

Birgit Kopainsky[1], Stephen M. Alessi[2], Pål I. Davidsen[1]

[1] System Dynamics Group, Department of Geography, University of Bergen, Postbox 7800, 5020 Bergen, Norway

[2] College of Education, University of Iowa, 370 Lindquist Center, Iowa City, IA 52242, United States

Abstract

When evaluating the effectiveness of interactive learning environments it is important to include measures of knowledge acquisition that complement measures of performance. In this paper we report on participants' knowledge acquisition in a dynamic decision making task where participants learned about and managed a small developing nation. In the course of the experiment participants not only had to make decisions but also answer multiple-choice questions and short essay questions. The results suggest that participants had a fairly good understanding of the reinforcing nature of national development processes and of processes that are in close causal proximity to their decisions. On the other hand, participants largely failed to recognize nonlinearities, the existence of the outflows to stocks and the proper treatment of delays with different durations. Knowledge acquisition was facilitated by the intensity of participants' exploration activities during a simulation-based, guided exploration phase between reading textual instructions and making actual, simulation-based decisions.

Introduction and background

One of the primary purposes of system dynamics has always been learning about complex dynamic systems (Forrester, 1961; Sterman, 2000). In just about all system dynamics activities, including when professionals are engaged in model building, learning is at least one of the important outcomes. It may not be the same type of learning as occurs in a classroom. But when researchers are working with new models, the knowledge derived from the research is a form of learning. When policy makers are testing the effects of new policies through modeling, insights into policy outcomes is also a form of learning. Hence, learning (the creation or acquisition of new knowledge) is always *one* of the purposes in system dynamics activities.

Given that assertion, that learning is always one of the purposes in system dynamics activities, then the accurate measurement of learning (or knowledge acquisition) is critical to advancement of system dynamics and the everyday practice (whether in a classroom, a lab, or a government office) of system dynamics activities. Measurement of learning is, however, quite varied. That is in part because the visible evidence, or artifacts, of learning are themselves varied. Skilled performance (solving a problem, designing a factory, running a business) is evidence of learning. Explaining or describing something (how to solve a problem, design a factory, or run a business) is also evidence of learning. A score on an exam is another type of evidence.

In our work with system-dynamics-based interactive learning environments (ILE) (Alessi, Kopainsky, Davidsen, & Pedercini, 2008; Kopainsky, Alessi, Pedercini, & Davidsen, 2009; Kopainsky, Pedercini, Alessi, & Davidsen, 2010; Kopainsky, Pirnay-Dummer, & Alessi, 2010), the assessment of learning outcomes is obviously a primary consideration. In this paper, using data from recent experimental studies, we examine different ways to measure different types of learning.

Within the fields of learning theory and learning assessment, one of the most widely used taxonomies for different aspects of learning is Bloom's Taxonomy of Educational Objectives (Bloom, 1956). The current version of this taxonomy (Anderson & Krathwohl, 2001) sequences six broad categories of learning outcomes from those representing lower level skills to those representing higher level skills. They are, starting at the lowest level:

1. Remembering. Remembering is represented (for example, on exams) by defining terms, naming objects, arranging things in order, repeating statements, and so on.
2. Understanding. Understanding is represented by being able to explain concepts, classify things into categories, describe principles, or restate information in one's own words.
3. Applying. Applying is represented by using knowledge and rules to solve problems, writing a report, or doing a job.
4. Analyzing. Analyzing is represented by comparing and contrasting, criticizing, asking good questions, generating hypotheses, or categorizing things.
5. Evaluating. Evaluating is represented by making judgments, arguing and debating, defending a point of view, or making an assessment.

6. Creating. Creating is represented by designing, planning, writing proposals, setting up experiments, drawing diagrams, building devices, and so on.

In any particular activity, the levels of learning that are intended should inform the types of assessment (or measurement) employed. Most activities have multiple levels of learning as their goals, though not all of the above levels.

Bloom's Taxonomy has already been applied to the design of instruments for measuring learning in systems dynamics by Stave & Hopper, (2007) and Hopper & Stave, (2008). They have investigated students' knowledge about systems thinking, which has been alternatively argued as either a subset or superset of the system dynamics methodology. That argument is not important to the research reported here, because we are measuring the acquisition of knowledge about specific content within a system-dynamics-based learning environment, not systems thinking in general or even system dynamics methodology in general. However, we consider Stave & Hopper's work as justification for using Bloom's Taxonomy to distinguish and measure different aspects of learning within learning environments based on either systems thinking or the system dynamics methodology.

Many system dynamics activities (model building, gaming, using a learning environment) focus on the participants' *performance* within the activity itself (the quality of the model created, whether they win the game, their score in the learning environment). While that is useful, it is not a complete picture for several reasons. First, performance is usually an indication of the *applying* level of Bloom's Taxonomy. Since applying knowledge is the third of six levels, it is entirely possible that a person might not perform well (i.e., apply what knowledge they have attained), but still might have demonstrated learning in one of the lower and easier levels, e.g., might have been able to show understanding or recall of the knowledge. Second, someone may display good performance for reasons other than learning, or as some in the system dynamics community like to say, they may perform well but for the wrong reasons. Common reasons for performing well without having learned the relevant knowledge include luck, intelligent guessing, trial and error, and cheating. We of course want people to perform well for the *right* reasons, because they understand models and how to apply them to real-world problems and organizations. Third, and this really overlaps with the first and second reasons, a researcher does not want to miss anything important. Learning at the lower levels is easy to attain and is common. Successively higher levels of learning are more difficult to attain and less common. It is therefore prudent to measure learning at several levels, not only the highest level you desire, but also those below it. If you only measure at the highest levels (of the taxonomy) that you desire, you may think nothing was learned when in fact people may have learned a lot, at the levels of remembering or understanding, for example. If you measure at several levels, you will know if only the most basic levels of learning occurred, but also can discern if higher levels of learning occurred, which would certainly be better.

For the above reasons, our research on learning environments includes measurement of both performance and what we have been calling (in previous papers and articles) understanding. In actuality, our measurement of "understanding" has been the measurement of a combination of remembering, understanding, applying, and analyzing, the first four levels of Bloom's Taxonomy. In this paper we analyze and discuss data concerning our different methods of measuring those aspects of learning. The highest levels

of Bloom's Taxonomy, evaluating and creating, are generally more relevant when people are building or modifying models, not playing games or running learning environments. For this reason they are not a part of this discussion.

This research on measurement of knowledge acquisition is in the context of our work on a learning environment called BLEND, the Bergen Learning Environment for National Development. It is intended to introduce national planners to systems thinking within their field. That includes appreciation for long term planning, feedback loops, non-linear relationships between variables, and collaborative planning across areas of an organization. However, our learning environments are intended for novices to system dynamics, that is, persons with little or no experience in it. The activities within the learning environment and our research measurements of knowledge acquisition probe the participants' knowledge about the specific structure underlying the system (e.g., the interactions between a nation's economy, social services, and infrastructure) and their decision making tasks to manage the system (e.g., allocating government funds to health, education, and transportation infrastructure). We do not use terms like "causal loop" or "stock and flow" in the learning activities, nor do we ask the participants to create such system dynamics artifacts. However, one of our long-term learning goals is to increase their (national planners) interest in the concepts and methodology of system dynamics.

In the version of BLEND used for the study reported here, participants played the role of adviser to the Prime Minister of a small developing nation, making long term decisions about investments in health, education, and transportation infrastructure (roads). Performance within the learning environment is fairly easily measured by computing the per capita income attained, adjusting for interest payments on the national debt. But as discussed above, it is possible for participants to succeed in their task through luck, good educated guessing, trial and error, or other means. We want our participants to manage their nation well because they understand and follow a good economic and social model, and because they exercise a systems approach. Thus, we must assess not only their performance, but their knowledge. In an earlier study (Kopainsky et al., 2009) we assessed knowledge using open-ended short essay questions, with moderate success. Although the results of that study showed differences in knowledge acquisition for different experimental conditions, there was very wide variation among participants. In this study we modified those short essay questions to make them more focused and we added an objective multiple-choice test. Well designed objective tests tend to have greater reliability and precision than subjective tests like short essays, plus they are easier to score.

All participants in this study answered both the objective (multiple choice) test and the subjective (short essay) story questions. However, half of the participants received a simulation interface which was higher in transparency (that is, made the underlying model more obvious to them) and half received a simulation interface which was lower in transparency (that is, made the underlying model less obvious, or more opaque). Both experimental conditions were based on an instructional strategy that we had developed in previous papers and that we termed prior exploration (Kopainsky & Sawicka, 2011; Kopainsky et al., 2009). This strategy includes a simulation-based, guided exploration phase between reading textual instructions and making actual, simulation-based management decisions.

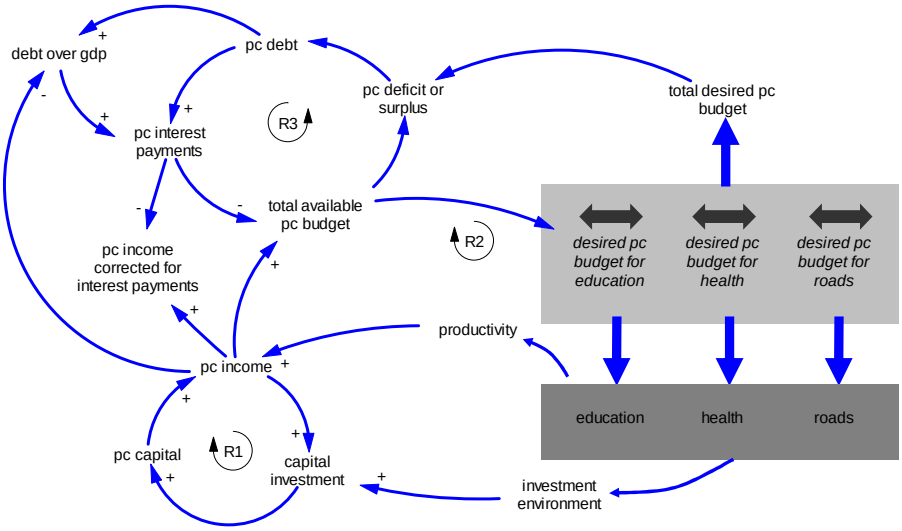
In this paper we use the experimental data to measure participants' knowledge acquisition and determine for which aspects of the national development planning tasks it was high and for which aspects it was lower. We also investigate how knowledge acquisition was affected by modifications in the prior exploration strategy and how the different activities in the task contributed to knowledge acquisition. We conclude the paper with reflections on the usefulness of the applied assessments to measure knowledge acquisition in a complex dynamic decision making task and with reflections on future research to improve such assessments.

Materials and methods

National development planning task

Participants played the role of the advisor to the prime minister in Blendia, a virtual sub-Saharan African Nation which, at the outset, is one of the poorest nations in the world (per capita income of $300 per person per year). Their task was to achieve and maintain the highest possible per capita income in the course of 50 years (see appendix A for the complete instructions). The prime minister has complete authority to decide on expenditures for education, health and roads. Investment and borrowing decisions are made every five years. The simulation starts in equilibrium and the prime minister stays in office throughout the 50 years no matter how poor a participant's performance. Figure 1 provides a simplified representation of the simulation model underlying the national development planning task. The sliders and the three variables in *italics* in the light grey box represent the three decision variables. The simulation model is described in detail in Kopainsky et al., (2009).

Figure 1: Simplified representation of the simulation model underlying the national development planning task



Notes:

R1: private sector development

R2: human and infrastructure development

R3: debt trap

Experimental conditions

Participants completed the national development planning task following the prior exploration strategy. The prior exploration strategy allows participants to explore the model's behavior in a risk-free environment after reading instructions and before making actual decisions (the management phase). Participants were split into two groups (the opaque and transparent groups) that received different versions of the exploration and management interface. The interfaces differed in the transparency of the underlying simulation model. The opaque group received an interface that showed sliders for the decision variables and six output graphs for key indicators (Figure 2). The interface of the transparent group showed the same sliders and output graphs but it linked the sliders and output graphs in the form of an aggregated causal loop diagram (Figure 3). In a separate paper we discuss the two experimental conditions in detail. In this paper we focus on the elements of the national development planning task for which knowledge acquisition was high, those for which it was low and on the determinants of knowledge acquisition.

Figure 2: Black box interface for exploration phase, step 1

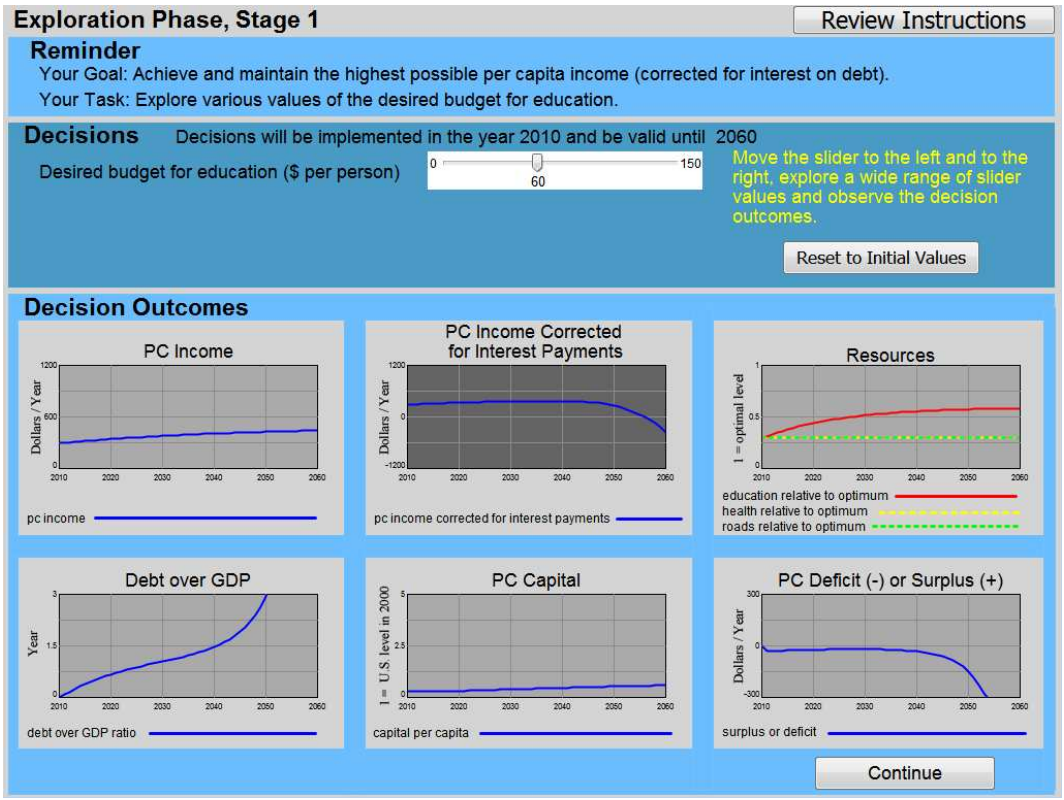
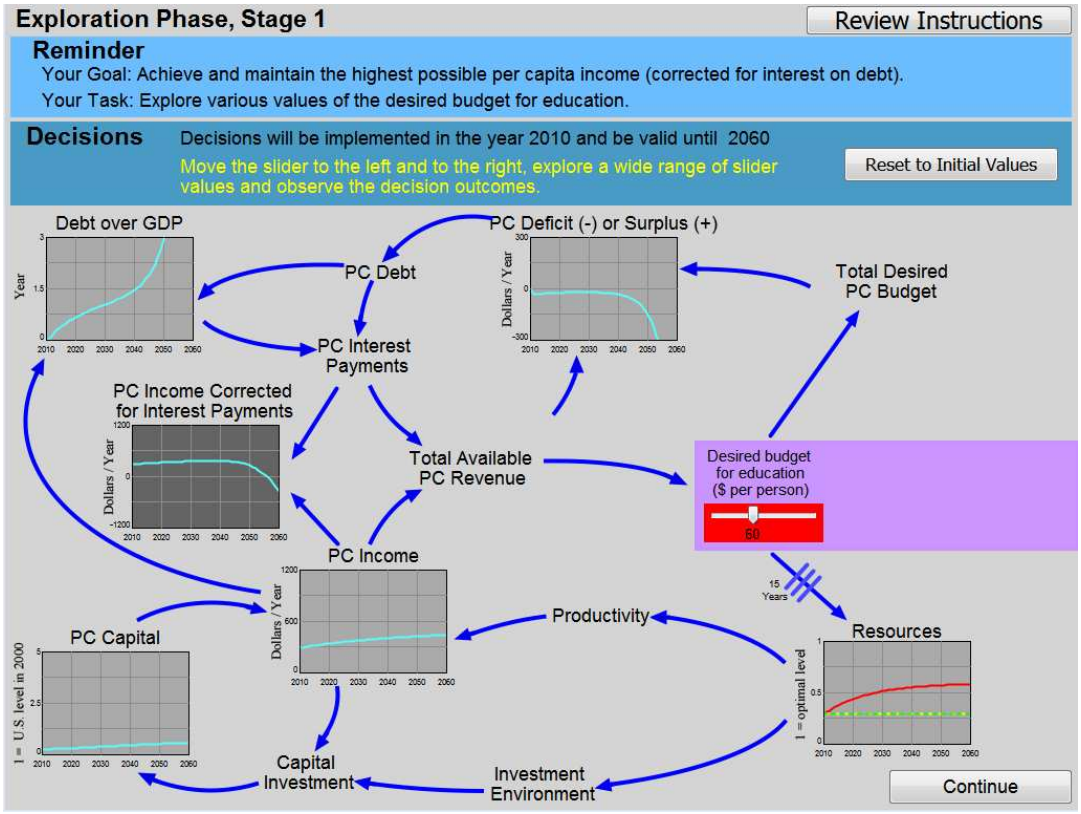


Figure 3: Transparent interface for exploration phase, step 1



Measures of knowledge acquisition

Objective multiple-choice test

The objective multiple-choice test consisted of eight questions (see Appendix B) each having six response alternatives. Four of the questions were designed to assess participants' recall and understanding (the bottom two levels of Bloom's Taxonomy) of the model underlying the nation's economic and social progress, and four of the questions were designed to assess participants' application and analysis (the next two higher levels of Bloom's Taxonomy) of the model and system dynamics principles. The test was administered twice, before participants started the simulation activities (the pretest) and again after completing all the simulation activities (the posttest). The order of the eight items and the order of the response alternatives was different (randomly) in the pretest and posttest.

Table 1 provides an overview of the questions in the pre- and the posttest and their correct answers. The table also lists the question identifiers (i.e., their short description that will be used in the results section of this paper) and the levels in Bloom's taxonomy that are assessed with the questions.

Table 1: Multiple-choice questions for pre- and posttest

Question identifier	Question stem wording	Correct answer	Level in Bloom's taxonomy
decisions in the task	The Prime Minister of Blendia can influence the following aspects directly	Expenditures for education, health, and roads	1, 2
determinants of tax rate	In the country of Blendia the tax rate	is fixed	1, 2
determinants of capital investments	In the country of Blendia, capital investment depends on:	The levels of education, health and roads	1, 2
determinants of interest rate	What determines the interest rate in Blendia?	The amount of debt and the GDP (pc income).	3, 4
mechanisms that lead to a decrease in debt	How can you pay down (service) debt in Blendia?	By distributing less than the total revenue.	3, 4
determinants of per capita income	In Blendia, economic development is measured by per capita income.	Per capita income in Blendia is the value of production per person and production is determined by the amount of physical capital, human capital and roads.	1, 2
length of delays	In the country of Blendia, which of the investments has/will have the most immediate effect on per capita income? Rank the resources and list the resource with the most immediate effect first.	Roads, health, education.	3, 4
mechanisms that lead to an increase in debt	High levels of debt in Blendia are a consequence of:	Spending more than earning through tax revenue.	3, 4

Subjective short essay questions

The subjective measure consisted of embedded story questions (Pirnay-Dummer, 2006). The participants were asked to imagine they are writing their advice to the Prime Minister. More specifically, they were asked to provide the Prime Minister with a verbal description of two things (see appendix C and D for the full wording of the questions):

- The problem situation in Blendia at the beginning of their assignment as the main advisor to the Prime Minister. This included identifying the key variables that are relevant to the problem and explaining the relationship between them.
- Their proposed strategy to increase per capita income while maintaining low interest payments on debt. This included explaining which policies they suggested implementing and why they thought these would be effective.

The embedded story questions were administered after the participants had completed all the simulation activities. The verbal protocols (participant responses) resulting from the embedded story questions were coded and assessed from three perspectives. In each case, descriptions of the problem situation and of the proposed strategy to solve the na-

tional development planning task were combined into one verbal protocol which was then compared to an expert response. The expert response also described the problem structure (i.e., the model structure shown in Figure 1) and the strategies for successfully solving the national development planning task. The three perspectives on the verbal protocols were:

- Coding and rating for detail complexity (manual analysis).
- Coding and rating for dynamic complexity (manual analysis).
- Automated assessment of structural and semantic similarity to the expert response.

For the manual analysis, the participants' responses were printed on one side of an index card and their participant identification number and experimental condition was on the reverse side to enable blind scoring. A scoring protocol was devised to assess participants' understanding of detail complexity and dynamic complexity (Senge, 1990). The scoring protocol awarded varying points to these elements with the maximum number of points determined by the expert response. In the expert response, we identified 16 relationships between important variables that served as indicators for *detail complexity*. An example of such a relationship is that per capita income depends on capital and total factor productivity. Participants received one point for each of those relationships that they identified, the maximum being 16.

To measure participants' understanding of *dynamic complexity*, points were assigned if participants were able to infer the characteristics of successful investment strategies. In total we coded the verbal protocols for a maximum of six such characteristics. Participants received one point if their description included the concept of balancing education, health and roads (recognition of nonlinearities, i.e., the fact that neither resource alone can stimulate growth very much but that they induce growth very effectively when developed in a balanced way). They received one point *each* if their description included education and roads requiring early investment and health requiring a somewhat delayed investment (recognition of stock variables with different delays in their inflows). Finally, they received one point *each* if they included the notion that borrowing early was important and that, at a later time, debt must be paid off (recognition of stock and flow variables, and understanding how these variables interact to produce an increase or a decrease in the stock). The scoring was fairly liberal, that is, any phrase suggesting they understood these key concepts was awarded a point.

Finally, the verbal protocols were also subjected to an *automated analysis* that we have introduced in one of our previous papers (Kopainsky, Pirnay-Dummer et al., 2010). The automated analysis was based on T-MITOCAR, a software tool that uses natural language expressions (instead of graphical drawings by participants) as input data for the re-representation, analysis and comparison of mental models (Pirnay-Dummer & Spector, 2008; Pirnay-Dummer & Ifenthaler, 2010). Such natural language expressions are the responses written by our participants in answer to the embedded story questions.

Any text of sufficient length can be graphically visualized by the T-MITOCAR software. T-MITOCAR tracks the association of concepts from a text directly to a graph, using mental model heuristics to do so. Texts which contain 350 or more words can be used to generate associative networks as graphs from text and to calculate structural and semantic measures for the analysis and comparison of mental models. The re-representation process is carried out automatically in multiple computer linguistic stag-

es. Table 2 provides an overview and definitions for the similarity indices calculated by T-MITOCAR. More details about the indices can be found in Kopainsky, Pirnay-Dummer et al., (2010).

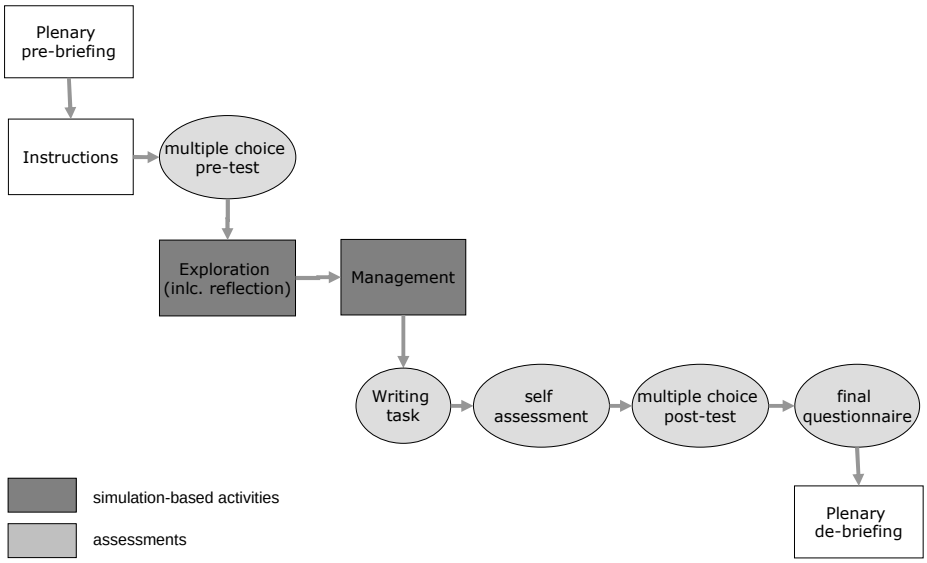
Table 2: Structural and semantic similarity indices used for the quantitative comparison of participant responses and an expert response

	Similarity index	Definition
Structure	*surface* measure (see Ifenthaler, 2008)	compares the number of link within two graphs. It is a simple and easy way to calculate how large a text model is.
	graphical matching measure (see Ifenthaler, 2008)	compares structural ranges of two graphs. It is calculated as the similarity between the diameters of the two spanning trees. The diameter of the spanning tree of a graph is the longest of the shortest paths between two (indirectly) linked concepts in a graph.
	density of vertices measure (also often called "*gamma matching* measure") (Pirnay-Dummer, Ifenthaler, & Spector, 2010)	describes the quotient of concepts per links within a graph. Since both graphs which connect every concept with all the other concepts (everything with everything) and graphs which only connect pairs of concepts can be considered weak mental models, a medium density is expected for most good working mental models.
	structural matching measure (see Pirnay-Dummer & Ifenthaler, 2010)	compares the complete structures of two graphs without regard to their content. This measure is necessary for all hypotheses which make assumptions about general features of structure (e.g., assumptions stating that expert knowledge is structured differently from novice knowledge).
Semantics	*concept matching* measure (Pirnay-Dummer et al., 2010)	counts how many concepts are alike. This measure is especially important for different groups operating in the same domain (e.g., using the same textbook). It determines differences in language use between the models.
	propositional matching measure (see Ifenthaler, 2008)	compares only fully identical propositions (concept-link-concept) between two graphs. It is a measure for quantifying semantic similarity between two graphs.
	balanced semantic matching measure (see Pirnay-Dummer & Ifenthaler, 2010)	a measure which combines both propositional matching and concept matching.

Procedures

Participants performed the experiment in a classroom setting with an experimenter always present in the room. Prior to the actual experiment the ILE was installed on the classroom computers. All results were stored electronically both on the desktop of each computer and to a remote server. Participants were assigned randomly to one of the two conditions. Before starting the task, all participants received the same pre-briefing. Pre-briefing emphasized that the participants were about to manage a virtual nation over a very long time horizon. They were then presented with the general schedule of the experiment (Figure 4).

Figure 4: Summary of the experimental procedure



The participants proceeded at their own pace and required between 45 and 90 minutes to complete all the activities. They worked at separate computers with no communication allowed with other participants. Performance measurement was based on the management task. The experimental session ended with a plenary debriefing session which included an exchange of participants' experiences while performing the experiment, collaborative development of the underlying model structure and a discussion of the short and long term effectiveness of different expenditure strategies.

In addition to the simulation-based activities, participants completed several questionnaires designed to explain their performance and assess their knowledge acquisition. After the participants had been introduced to the nation of Blendia (the instructions), participants took the multiple-choice pretest.

The exploration phase consisted of four steps. In step one, participants could manipulate the slider regulating education expenditure and observe the resulting model behavior. In step two, they could do the same for health expenditure, and in step three, for roads expenditure. In the fourth and last step, they could manipulate all three sliders at the same time. Exploration is a dynamic activity. As the participant slides the slider for education, health and/or roads higher and lower, the graphs in the interface immediately replot to show how the selected budget would affect the various outcome variables. After each exploration step, participants were asked to record their observations and to explain the resulting graph behavior. These questions, and the reflection they were intended to encourage, are considered an integral part of the prior exploration strategy, and not as measurement of outcomes.

In the management phase, participants had to make and implement decisions about the expenditures for education, health and roads. During the management phase, moving a slider does not immediately affect the graphs for the outcome variables. Only when the participant clicks the button labeled "Click Here to Simulate for the Next 5 Years" do the graphs update to show the outcomes for that 5-year period. The participant can then move the sliders again to modify the investment strategy. This process (modify the sliders, go forward 5 years) is done ten times.

After the management task the participants answered the short essay questions, that is, the embedded story questions. In addition to the short essay questions, participants also assessed the usefulness of their proposed strategy and the usefulness of the simulation in nine questions on a five-point Likert scale ranging from “strongly disagree” to “strongly agree” (appendix E). After the self-assessments, participants were administered the multiple-choice posttest. In a final questionnaire, participants were asked to assess their interest in, prior knowledge of, and experience with national development issues and the use of simulations for national planning. Participants also indicated their highest degree of education, their age category and their gender. This was the participants’ background and demographic data and was collected so that analysis could control for the effects of their backgrounds on knowledge acquisition (Appendix F).

Participants

Data was collected with 39 introductory level system dynamics students in the fall 2010. The students were recruited from the University of Bergen in Norway and ETH Zurich in Switzerland. Participants were assigned randomly to either the opaque (21 participants) or transparent group (18 participants). Participants were 46% female and 54% male. 82% were between 22 and 30 years old and 18% were above 30.

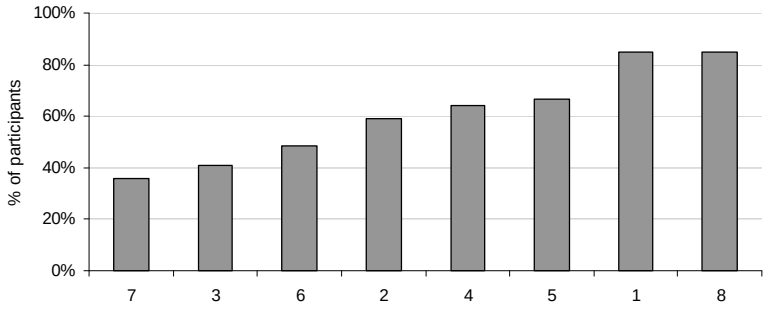
Results

The results presented in the subsequent sections did not differ significantly between the two experimental conditions. Those included Mann-Whitney tests at $\alpha=0.05$ and tested for differences between conditions on performance, multiple-choice pretest, multiple-choice posttest, number of described relationships in the short essay questions, number of described strategy characteristics in the short essay questions, similarity indices calculated by T-MITOCAR for the short essay questions, range of slider movements during the four exploration steps, time spent on the four exploration steps, self assessment of strategy and simulation, demographic and background information. For this reason we present results aggregated over the two conditions and focus on identifying those aspects of the national development planning tasks that participants understood well as well as identifying aspects with which they had more difficulty. In a last step we analyze which activities in the task were useful for explaining participants’ knowledge acquisition.

Multiple-choice questions

Figure 5 provides an overview of the results from the multiple-choice questions in the posttest. For each question the percentage of participants who answered it correctly is indicated. The questions are arranged in order of increasing percentage of correct answers.

Figure 5: Correct answers to the multiple-choice questions in the posttest

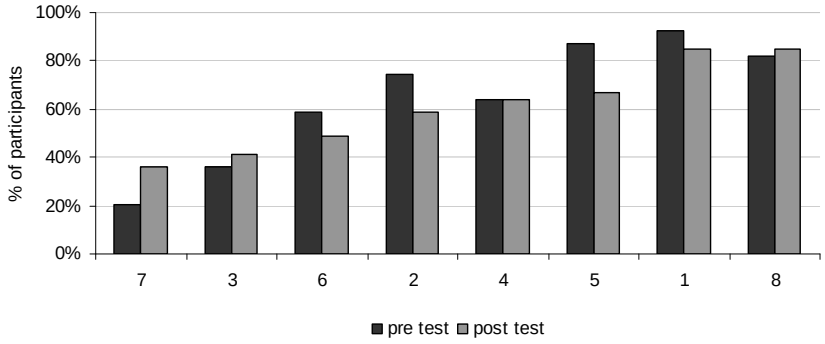


1: length of delays
2: determinants of tax rate
3: determinants of interest rate
4: mechanisms that lead to an increase in debt
5: determinants of capital investment
6: determinants of per capita income
7: mechanisms that lead to a decrease in debt
8: decisions in the task

Figure 5 indicates a high percentage of participants correctly answered the questions about the decisions in the task (question 8) and about the length of the delays for education, health and roads (question 1). Questions about relationships close to the decisions in the task (questions 2 and 5) and the mechanisms that lead to an increase in debt (question 4) were also answered correctly by around 60% of the participants. Relationships close to the decisions in the task relate to relationships that are either a direct consequence or a direct determinant of the three decision variables (see Figure 1). A small percentage of participants were able to correctly answer questions about distant relationships (question 3) and the requirements for reducing the debt stock (question 7). The question about the determinants of per capita income (question 6) achieved a low-to-mid-range percentage of correct answers.

The results in the pretest were similar to the results in the posttest. Figure 6 compares the percentage of participants who answered the questions in the tests correctly. The figure illustrates that there was a tendency towards a lower percentage of correct answers in the posttest compared to the pretest. The difference between the two tests was significant (Wilcoxon test at $\alpha = 0.05$) for questions 2 and 5. These questions were fairly simple questions that required participants to recall that they could not change the tax rate in the simulation (question 2) and that capital investment depends on the expenditures for education, health and roads (question 5).

Figure 6: Correct answers to the multiple-choice questions in the pre- vs the posttest



1: length of delays

2: determinants of tax rate

3: determinants of interest rate

4: mechanisms that lead to an increase in debt

5: determinants of capital investment

6: determinants of per capita income

7: mechanisms that lead to a decrease in debt

8: decisions in the task

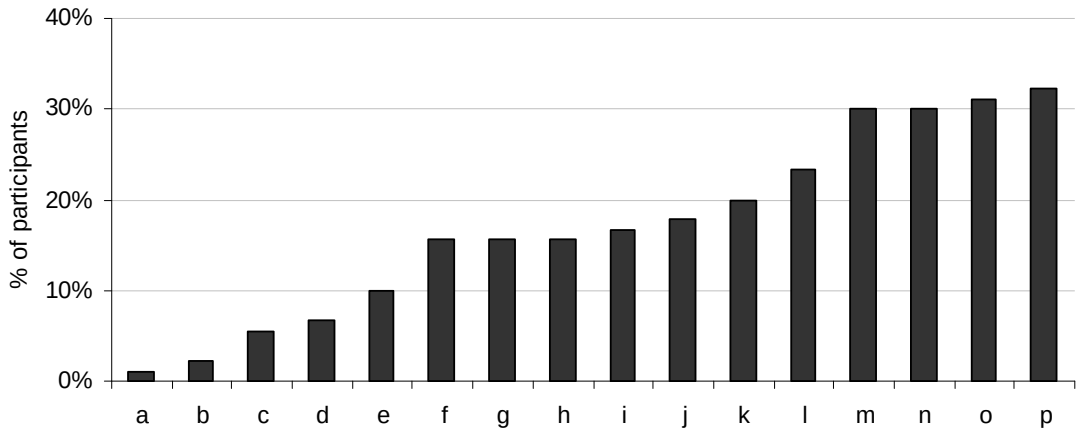
Embedded story questions

Figure 7 provides an overview of the relationships described in the verbal protocols (detail complexity). The relationships are arranged in order of an increasing percentage of participants who described them in their answers to the embedded story questions. Figure 7 illustrates that those relationships that participants described most were the goal of and the decisions in the task (relationships l and m) as well as relationships that are very close to the actual decisions (relationships n, o and p).

Poorly described relationships were the mechanisms that lead to a decrease in debt (relationships a, and e) and relationships that are distant from the decisions in the task (relationship b). The low value for the relationship between capital and investment (relationship c) indicates that participants had difficulties perceiving the input to the capital stock.

The number of participants describing the determinants of capital investment (relationships g, h, and k) was in the medium range, and the number of participants describing the determinants of per capita income was a little smaller (low-to-medium range).

Figure 7: Level of detail complexity described in the verbal protocols



a surplus occurs when desired expenditures are less than available expenditures
b investment increases with per capita income
c capital increases with investment
d deficit occurs when desired expenditures are greater than available expenditures
e surplus leads to paying down debt
f per capita income is a function of capital and total factor productivity
g investment increases with education
h investment increases with health
i deficit leads to borrowing
j available expenditures are equal to tax revenue minus interest payments
k investment increases with roads
l goal is to maximize per capita income
m the prime minister can regulate expenditures
n tax revenue equals per capita income times the tax rate
o borrowing leads to debt
p debt leads to interest payments

Figure 8 shows the level of dynamic complexity described in the verbal protocols. The characteristics of a successful strategy for solving the national development planning task are arranged in order of increasing frequency that participants described them. The figure illustrates that half of the participants identified in their responses the necessity of borrowing early to kick start economic growth in Blendia. Only around 25% of the participants mentioned the need to develop the three resources education, health and roads in a balanced way (recognition of nonlinearities) and the need for paying down debt after the initial period of borrowing (suggesting a misperception of the outflow of the debt stock). Participants had particular difficulties with the management of the different delay times. This is reflected in the varying percentages of participants who mentioned the correct management of the delays (health later, education early, roads early).

Figure 8: Level of dynamic complexity described in the verbal protocols

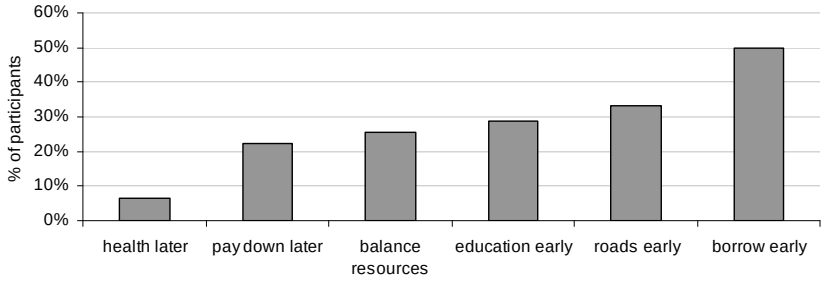
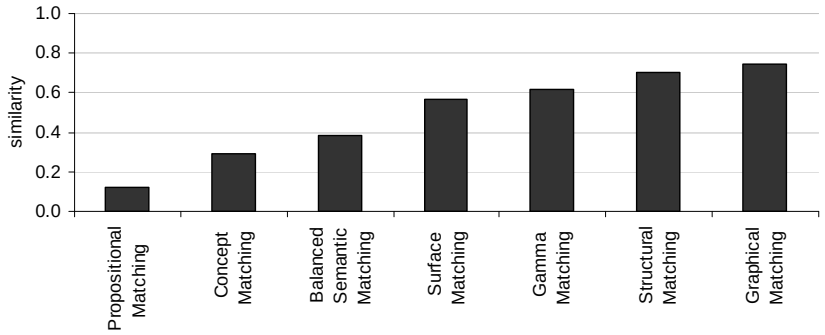


Figure 9 presents the results from the automated analysis of the verbal protocols using T-MITOCAR. The similarity indices in the figure indicate the overall similarity between the participants' responses and the expert response. A value of 1 for any of the indices in the figure would indicate that the participant response is identical to the expert response for a specific structural or semantic characteristic.

Figure 9: Structural and semantic similarity between the verbal protocols and an expert response



Results from the automated analysis show that in general, similarity between participants' responses and the expert response is considerably higher for the structural indices than for the semantic indices. Within the structural indices (graphical, structural, gamma, and surface matching), we can observe that participants do not describe very many concepts (variables) in their responses (fairly low level for surface matching), but that they link these concepts quite intensively (high levels for graphical and structural matching). The low values for concept and propositional matching, however, indicate that the few concepts that they describe are not very relevant in the national development planning task (low level of concept matching) and that they do not link the concepts correctly (low level of propositional matching).

Determinants of knowledge acquisition

After identifying those aspects of the national development planning task for which knowledge acquisition was high and those for which is was lower we analyzed how the activities in the task affected participants' knowledge acquisition. For this purpose we use regression analysis with the dependent variables being our knowledge acquisition measures and the independent variables being details about the activities in the national

development planning task and participants' background information. Figure 10 provides an overview of how much time participants devoted to the activities in the task.

Figure 10: Average percent of time spent on the activities in the national development planning task

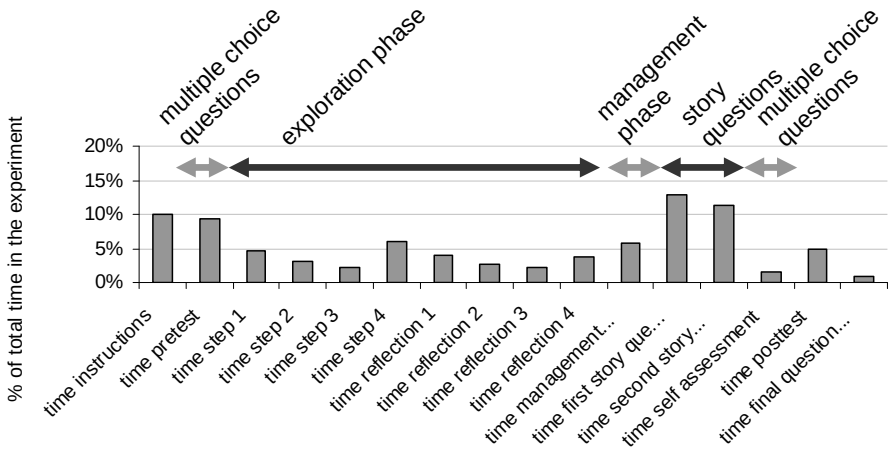
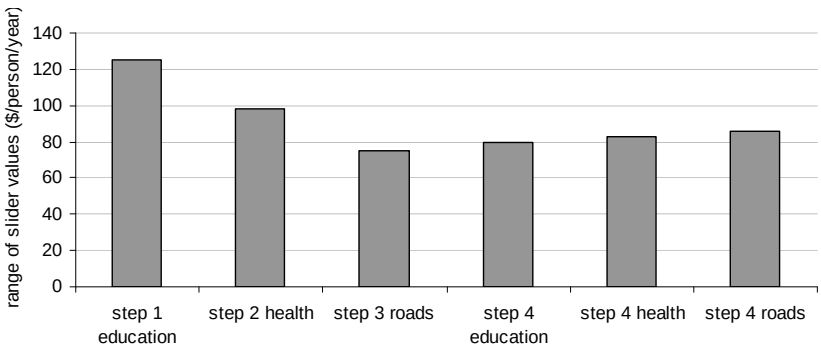


Figure 11 details the activities in the exploration phase. During exploration, participants were encouraged to explore a wide range of slider values for education, health and roads expenditures and observe the effect these changes had on the system's behavior. The figure shows the range of slider movements, i.e., the range of explored expenditures, for the three resources in the four exploration steps.

Figure 11: Range of slider movements in the four steps of the exploration phase



Other factors that may have influenced participants' knowledge acquisition include the participants' demographic characteristics and their background (data collected in the final questionnaire), which can be summarized as follows:

- Average interest in national development issues: 2.05 on a scale from 1 (very low interest) to 5 (very high interest).
- Average self assessed knowledge about national development issues: 2.92 on a scale from 1 (very poor knowledge) to 5 (very good knowledge).
- Proportion of participants who have taken classes in national development: 59%.
- Proportion of participants with experience in national development work: 10%.

- Proportion of participants who have used simulation to study national development issues: 13%.
- Highest education: high school 10%, bachelor 64%, master 26%.
- Age: 82% between 22 and 30, 18% above 30.
- Gender: 46% female, 54% male.

Using multiple linear regression we tested how the activities in the task and participants' characteristics affected knowledge acquisition. Table 3 summarizes the outcomes of three regression analyses with the dependent variables being (1) the total number of described relationships (embedded story question), (2) the total number of described strategy characteristics (embedded story question) and (3) the total correct answers in the multiple-choice posttest. Appendix G contains the complete output tables for the three models. Table 3 lists only the variables that were significant at α=0.05. For each significant variable the direction of its influence is indicated with either a "+" or a "-".

Table 3: Significant variables for explaining knowledge acquisition (significant at α=0.05)

	dependent variable: number of relationships	dependent variable: number of strategy characteristics	dependent variable: number of correct answers in multiple-choice posttest
time spent on exploration step 2			–
time spent on exploration step 4			+
time spent on first embedded story question		+	+
range of roads slider movements step 3		+	
age		+	

Table 3 shows that the regression model with total number of described relationships as the dependent variable did not identify any significant influencing factors that helped explain knowledge acquisition. The other two regression models overlapped in identifying the time spent on answering the first embedded story question as a significant influencing factor. Both models identified additional, albeit different, significant influencing factors. The time spent in the fourth exploration step (the step where participants could explore the model's reaction to changes in education, health and roads expenditure) had a significant and positive impact on the total number of correct answers in the posttest while the time spent on the second exploration step influenced the total number of correct answers negatively. The range of roads slider movements during exploration and age significantly contributed to the total number of strategy characteristics.

Discussion and conclusions

In this paper we used subjective and objective tests to assess learning in a system-dynamics-based interactive learning environment. The results from the objective multiple-choice tests and the subjective embedded story questions proved to be consistent. Participants in the national development planning task had a fairly good understanding of the decisions in the task and of the relationships in close causal proximity to those decisions. Knowledge acquisition for relationships distant to the decisions and for the mechanisms that decrease debt, on the other hand, was low. In both tests, knowledge

acquisition about the determinants of per capita income and of capital investment was in a medium range.

As could be expected, there was a tendency for participants to correctly answer those multiple-choice questions that assessed lower level skills and to have more difficulty with answering questions that assessed higher level skills. The levels refer to the six broad categories of learning outcomes in Bloom's taxonomy. However, we also found that the distance between the causal relationships and the task decisions was more important for explaining the percentage of correct answers than the actual level in Bloom's taxonomy. The multiple-choice tests also confirmed people's problems with stock management: the mechanisms that increase debt (i.e., the inflow) were understood better than the mechanisms that decrease debt (i.e., the outflow) (see Moxnes & Saysel, 2009; Booth Sweeney & Sterman, 2000; Sterman, 2010).

In addition to the two assessments showing consistent results, they complement each other in several ways:

- Participants' verbal protocols frequently *described* the mechanisms that lead to an increase in debt. However, a much smaller percentage of participants were able to correctly *predict* the outcome of these mechanisms (i.e., the frequency of correct answers for mechanisms that lead to an increase in debt in the multiple-choice tests was in the mid range).
- Verbal protocols provided a richer and more differentiated description of participants' understanding of the problem structure and effective solution strategies. Most participants described the education, health and roads stocks and fairly often they also described the debt stock. However, they rarely mentioned the capital stock. In their descriptions it was investment that influenced production and thus per capita income. Similarly, they often said that borrowing causes interest payments (instead of correctly tracing the causality from borrowing that increases the debt stock and that the amount of debt influences interest payments on debt). Also, participants only described how education, health and roads improve total factor productivity and thus per capita income. They rarely described how the three resources improve the investment environment which leads to increases in capital investment, capital and eventually production (per capita income). Participants realized that the three resources have different implementation delays (consistent with the results from the multiple-choice tests). However, they made the wrong conclusion by deciding to spend more on the resource with the shortest delay first instead of prioritizing the resource with the longest implementation delay.
- From the characteristics of the multiple-choice versus embedded story questions it could be expected that verbal protocols provided a richer and more differentiated description of participants' mental models. However, contrary to intuition or expectation, the number of described relationships does not seem to be a very good indicator for assessing the effects of instructional strategies on participants' knowledge acquisition. The number of described relationships did not generate valid regression models, neither for exploration related determinants nor for background related determinants of knowledge acquisition. The other two indicators (strategy elements, correct answers in the post test), on the other hand, did generate valid models and comparable results.

Overall, the regression models suggest that the intensity of exploration, measured in time spent on and the range of the exploration activity, positively contributed to knowledge acquisition. The fact that time spent on the second step of exploration negatively affected knowledge acquisition indicates a delicate balance between the importance of exploring the system's behavior and the possible danger of doing so excessively. During the first three steps of exploration, participants could only manipulate one expenditure category (education, health or roads) at a time. It seems that they first needed to pass these three steps without trying to understand everything at once before they were ready to appreciate the dynamics involved in the national development planning task in the fourth step.

From our results we derive the following conclusions for improving our learning environment and the experimental procedures:

- The number of correct answers in the posttest was lower than in the pretest. The reasons for this are unclear. Participants spent a much shorter amount of time on the multiple-choice questions in the posttest than on questions in the pretest (Figure 10). Comments from participants when they were performing the experiment indicated that they were confused about the fact that they had to answer the same questions again. It is also possible that participants thought that they needed to change their answers in a second round of multiple-choice questions. Finally, it is also conceivable that the multiple-choice questions in the pretest had an effect on knowledge acquisition. We clearly did not intend for the questions to have such effect. For all these reasons future versions of the ILE and the experimental design will omit the multiple-choice pretest.
- The current experimental design did not allow for construct validation as we could not measure any difference in knowledge acquisition between the two experimental conditions with our knowledge assessment tests. Our expectation was that learning would be greater in the transparency group than in the opaque group. Finding that result with the different measures of learning would have provided evidence of their validity with respect to measurement of knowledge acquisition. Future research will have to determine whether the lack of these findings is due to flaws in our measures or to the specific characteristics of the two experimental conditions (see also Kopainsky, Alessi, & Pirnay-Dummer, submitted).
- The general assumption is that better understanding of the structure underlying the decision making task and its resulting behavior will generate better performance, i.e., better decision making. With our experimental design it is, however, unclear whether performance in the decision making task might also influence knowledge acquisition. In future experiments we might therefore move the embedded story questions as well as the multiple-choice posttest immediately after the exploration phase but before the actual management phase.

References

Alessi, S. M., Kopainsky, B., Davidsen, P. I., & Pedercini, M. (2008). *A system dynamics-based multi-user domain for improving national development planning*. Paper presented at the Annual Meeting of the American Educational Research Association.

Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Boston: Allyn & Bacon.

Bloom, B. S. (1956). *Taxonomy of Educational Objectives: Book 1 – Cognitive Domain*. New York: Longman.

Booth Sweeney, L., & Sterman, J. D. (2000). Bathtub dynamics: initial results of a systems thinking inventory. *System Dynamics Review, 16*(4), 249-286.

Forrester, J. W. (1961). *Industrial dynamics*. Cambridge, MA: MIT Press.

Hopper, M., & Stave, K. A. (2008). *Assessing the effectiveness of systems thinking interventions in the classroom.* Paper presented at the 26th International Conference of the System Dynamics Society, Athens.

Ifenthaler, D. (2008). Relational, structural, and semantic analysis of graphical representations and concept maps. *Educational Technology Research and Development.*

Kopainsky, B., Alessi, S. M., Pedercini, M., & Davidsen, P. I. (2009, July 26-30, 2009). *Exploratory strategies for simulation-based learning about national development.* Paper presented at the 27th International Conference of the System Dynamics Society, Albuquerque, NM.

Kopainsky, B., Alessi, S. M., & Pirnay-Dummer, P. (submitted). *Providing structural transparency when explor-ing a model's behavior: Effects on performance and knowledge acquisition.* Paper presented at the 29th International Conference of the System Dynamics Society, Washington, DC.

Kopainsky, B., Pedercini, M., Alessi, S. M., & Davidsen, P. I. (2010). A blend of planning and learning: Simplifying a simulation model of national development. *Simulation & Gaming, 41*(5), 641-662.

Kopainsky, B., Pirnay-Dummer, P., & Alessi, S. M. (2010). *Automated assessment of learners' understanding in complex dynamic systems*. Paper presented at the 28th International Conference of the System Dynamics Society.

Kopainsky, B., & Sawicka, A. (2011). Simulator-supported descriptions of complex dynamic problems: Experimental results on task performance and system understanding. *System Dynamics Review*.

Moxnes, E., & Saysel, A. K. (2009). Misperceptions of global climate change: information policies. *Climatic Change, 93*(1), 15-37.

Pirnay-Dummer, P. (2006). *Expertise und Modellbildung: MITOCAR.*, Freiburg University, Freiburg.

Pirnay-Dummer, P., & Ifenthaler, D. (2010). Automated Knowledge Visualization and Assessment. In D. Ifenthaler, P. Pirnay-Dummer & N. M. Seel (Eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*. New York: Springer.

Pirnay-Dummer, P., Ifenthaler, D., & Spector, J. M. (2010). Highly integrated model assessment technology and tools. *Educational Technology Research and Development. , 58*(1), 3-18.

Pirnay-Dummer, P., & Spector, J. M. (2008). *Language, Association, and Model Re-Representation. How Features of Language and Human Association can be Utilized for Automated Knowledge Assessment.* Paper presented at the AERA 2008, TICL SIG, Chicago, Illinois.

Senge, P. M. (1990). *The fifth discipline: The art and practice of the learning organization*. New York: Doubleday.

Stave, K. A., & Hopper, M. (2007). *What constitutes systems thinking? A proposed taxonomy.* Paper presented at the 25th International Conference of the System Dynamics Society, Boston, MA.

Sterman, J. D. (2000). *Business dynamics. Systems thinking and modeling for a complex world*. Boston et. al.: Irwin McGraw-Hill.

Sterman, J. D. (2010). Does formal system dynamics training improve people's understanding of accumulation? *System Dynamics Review, 26*(4), 316-334.

Appendix

Appendix A: Instructions

You have just been appointed as the head advisor to the Prime Minister of Blendia. The Prime Minister and you will stay in office for a period of 50 years. You are thus in charge of the long term development of Blendia.

Blendia is an island located off the western coast of Africa. It is currently one of the poorest countries in the world with a per capita income of $300 per year. Your task is to bring the country onto a sustainable economic growth path and achieve and maintain the highest possible per capita income.

Per capita income results directly from production and sale of goods and services. For simplicity, assume that per capita GDP (per capita production) is equal to per capita income. Production is driven by the available physical capital (machinery and its technology level), by human capital (the amount of workers, and their education and health), and by the level of infrastructure (including roads). The government cannot invest in physical capital directly, but it can invest in improving the general level of education, health, and infrastructure. By investing in such resources, the general investment environment improves. Investors in capital will invest the potentially available money (a share of per capita income) more when the labor force is more productive and roads provide access to input and output markets for the goods produced.

Specifically, the Prime Minister can invest in the following three resources:

- Education

  Education is the stock of knowledge, skills, techniques, and capabilities embodied in labor acquired through education and training. These qualities are important for the labor force to understand and perform tasks, to properly use the available physical capital, and to efficiently organize the production process. Maximum or optimal education would mean an average adult literacy rate of 100%, which is the maximum or optimal value for Human Development Index (HDI) calculations. The HDI is a United Nations composite index that includes measures of education, health, and income. It allows comparison across countries of their level of human development.

- Health

  Health defines the strength of the labor force and thus its capability to properly use the available physical capital and to efficiently organize the production process. Maximum or optimal health would mean an average life expectancy of 85 years (which is the maximum or optimal value for Human Development Index calculations).

- Roads

  Efficient and extended infrastructure allows faster and cheaper access to the market, broader access to information, and reliable access to the inputs required for production. Maximum or optimal roads would mean a value of kilometers of roads per person equal to those in the year 2005 in the United States.

Budget issues

The budget for education, health and roads expenditures (also called "development expenditure") can be calculated as follows:

> + Revenue: Through taxation (30% flat tax rate) the government generates revenue from per capita income.

+ Borrowing: The government can borrow money from foreign sources (e.g., the International Monetary Fund). If the government borrows money, it starts accumulating debt.

- Interest payments on debt: Each year the government will have to pay interest on its debt. The interest rate depends on the level of debt. A common measure for the amount of debt is the debt over GDP ratio. The interest rate is 1% for a very low debt over GDP ratio and can rise up to 15% for a very high debt over GDP ratio.

Note that Revenue and Borrowing add funds (the plus signs) available for expenditures, while Interest payments on the debt subtract funds (the minus signs) available for expenditures.

Decisions

Every five years, as part of a national development planning effort, the Prime Minister will decide on the expenditures for education, health and roads. The Prime Minister can do three things, and has the absolute power to decide which to do (see also Figure 1):

1. Distribute the total available Per Capita Revenue among education, health and roads without creating either a deficit or a surplus.
2. Distribute more than the total available Per Capita Revenue. In this case the Prime Minister creates a deficit and borrows money.
3. Distribute less than the total available Per Capita Revenue. In this case the Prime Minister will have a surplus and be able to service (pay down) debt or lend money.

Figure 1: Budget decisions mechanism with initial values

Total available Per Capita Revenue	$90 per person
Education expenditure	$30 per person
Health expenditure	$30 per person
Transportation expenditure	$30 per person
Surplus (+) / deficit (-)	$0 per person

Evaluation

The performance of the Prime Minister will be evaluated based on a composite income indicator. The indicator is calculated as:

+ Per capita income: You should try to achieve and maintain the highest possible per capita income. The country's official goal is to reach a value of $600 per capita or more in 50 years.

- Interest payments on debt: Per capita income can only be maintained if the country has not accumulated excessive debt.

In summary, the interest payments on debt will be deducted from per capita income.

Appendix B: Multiple-choice questions

The same questions were used for the pretest and posttest. Questions and alternatives were presented in random order. The order here reflects the numbering of the questions in the main text. Correct answers are **highlighted**.

1. In the country of Blendia, which of the investments has or will have the most immediate effect on per capita income? Rank the resources, listing the resource with the most immediate effect first.

- Roads, education, health.
- **Roads, health, education.**
- All have their effect at the same time.
- Education, health, roads.
- Education, roads, health.
- Health, education, roads.

2. In the country of Blendia the tax rate

- **is fixed.**
- depends on the level of debt.
- is per capita income minus total expenditures.
- is tax revenue plus borrowing.
- is per capita income minus debt.
- depends on the total expenditures for education, health, and roads.

3. What determines the interest rate in Blendia?

- **The amount of debt and the GDP (per capita income).**
- GDP (per capita income) and the negotiation power of Blendia towards the lender country.
- How much Blendia is borrowing in the current year.
- How much Blendia borrowed the preceding year.
- The credibility that Blendia has due to its current amount of debt.
- The credibility that Blendia has due to its current amount of debt balanced by what it usually pays down.

4. High levels of debt in Blendia are a consequence of:

- Changing modalities in loan contracts.
- **Spending more than earning through tax revenue.**
- Mismanagement and corruption by government officials in Blendia.
- The geographic disadvantages of Blendia.
- The lack of natural resources in Blendia.

- Budged shortages with donor agencies.

5. In the country of Blendia, capital investment depends on:

- The total government development expenditure.
- The government's expenditures on education, health and roads.
- **The levels of education, health and roads.**
- The tax revenue minus interest payments on debt.
- The tax rate minus the interest rate.
- The level of education and the tax revenue minus the interest payments on debt.

6. In Blendia, economic development is measured by per capita income. Per capita income in Blendia is the:

- value of production per person and production is determined by the amount of physical capital minus interest payments on debt.
- sum of the government's expenditures on education, health and roads per person.
- sum of the government's expenditures on education, health and roads per person minus interest payments on debt.
- **value of production per person and production is determined by the amount of physical capital, human capital and roads.**
- sum of tax revenue and borrowing minus interest payments on debt.
- tax revenue minus the sum of the government's expenditures on education, health and roads per person.

7. How can you pay down (service) debt in Blendia?

- By borrowing more money from foreign sources.
- **By spending less than the total revenue.**
- By spending more than the total revenue.
- By negotiating debt relief.
- By raising taxes for a short period of time.
- By raising taxes for a long period of time.

8. The Prime Minister of Blendia can influence the following aspects directly:

- **Expenditures for education, health, and roads.**
- Level of debt, capital investment, and tax rate.
- Expenditures for roads, tax rate, and capital investment.
- Expenditures for education, health, and level of debt.
- Interest rate (on debt), tax rate, and capital investment.
- Expenditures for roads, level of debt, and interest rate (on debt).

Appendix C: Embedded story question – part I

As the Prime Minister's main advisor, you must clearly understand the situation in Blendia and steps necessary to achieve and maintain the highest possible per capita income. The Prime Minister will be traveling to an important United Nations conference where heads of sub-Saharan African nations will meet to discuss strategies for breaking out of the poverty trap. The country with the best strategy will receive the most favorable loan conditions from the International Monetary Fund.

On this and the next page you will prepare a concept note for the Prime Minister, explaining in detail why Blendia has such a low per capita income and what the Prime Minister must do to change this, i.e., how much money the Prime Minister must spend on education, health and roads every five years throughout the next 50 years. Bear in mind that the Prime Minister is a politician who does not have much time to think about the causes of poverty and why your strategies would succeed. You must explain yourself very clearly and include as much relevant information as possible.

In the spaces below, describe Blendia's problem situation to the Prime Minister. Try to identify the key issues or variables relevant to the problem and explain the relationship between them. Please give the Prime Minister your six most important ideas in enough detail that the Minister will clearly understand what you are saying.

Appendix D: Embedded story question – part II

Now, in the space below, explain for the Prime Minister your insights and suggestions about increasing per capita income in Blendia while maintaining low interest payments on debt. How much money should the Minster spend on education, health and roads over the next 50 years? Be as specific as possible and explain the reasons for each step in your strategy. This is important because the Prime Minister must be able to give a very convincing rationale to other Ministers at the conference.

Appendix E: Self assessment questions

Please give us your opinion on the following statements. Click on the diamonds.

	I strongly disagree	I disagree	I neither dis-agree nor agree	I agree	I strongly agree
My proposed strategy will definitely help Blendia if it is implemented.					
I am sure that the Prime Minister will understand my strategy.					
I am sure that the Prime Minister will implement my suggestions.					
I think that my suggestions are easy to implement.					
I believe that the people of Blendia will understand my strategy.					
The simulation helped me to create a good strategy.					
The simulation made a lot of things clear to me.					
Running the simulation has influenced my ideas about the problem in Blendia.					
Running the simulation has positively influenced my interest in the field.					

Appendix F: Final questionnaire

How interested are you in national development issues?

- Extremely
- Quite
- Some
- Not particularly
- Not at all

Have you ever taken classes in national development studies or in national development economics?

- Yes
- No

Have you ever used simulation and modeling to study or manage national development issues?

- Yes
- No

What is your age?

- Below 18 years
- 18 to 21 years
- 22 to 30 years
- Above 30 years

How would you rate your knowledge of national development issues?

- Very good
- Good
- Average
- Poor
- Very poor

Do you have any practical experience in national development work?

- Yes
- No

What is your highest educational degree?

- Secondary School
- B.A.
- M.A.
- Ph.D.

What is your gender?

- Female
- Male

Appendix G: Regression models for explaining knowledge acquisition

Dependent variable: total number of described relationships in the embedded story questions

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	Kollinearitätsstatistik	
		RegressionskoeffizientB	Standardfehler	Beta			Toleranz	VIF
1	(Konstante)	-.416	1.209		-.344	.735		
	EXPL_quant_edu_step1	-.005	.005	-.283	-1.003	.329	.260	3.842
	EXPL_quant_health_step2	.004	.006	.152	.622	.541	.346	2.889
	EXPL_quant_roads_step3	.006	.007	.207	.864	.399	.362	2.765
	EXPL_qual_step_1_0_no_explanation_1_wrong_explanation_of_behav	-.174	.501	-.073	-.346	.733	.469	2.133
	EXPL_qual_step_2_0_no_explanation_1_wrong_explanation_of_behav	-1.278	.969	-.527	-1.319	.203	.130	7.702
	EXPL_qual_step_3_0_no_explanation_1_wrong_explanation_of_behav	1.035	1.000	.459	1.036	.313	.105	9.495
	EXPL_qual_step_4_0_no_explanation_1_wrong_explanation_of_behav	1.422	.485	.640	2.934	.009	.435	2.297
	TIME_EXPL_exploration_step_1	.004	.004	.369	1.169	.257	.208	4.805
	TIME_EXPL_exploration_step_2	-.003	.006	-.175	-.513	.614	.178	5.630
	TIME_EXPL_exploration_step_3	.006	.007	.226	.822	.421	.275	3.641
	TIME_EXPL_exploration_step_4	7.545E-5	.000	.033	.181	.858	.608	1.646
	TIME_EXPL_reflection_1	-.001	.003	-.109	-.409	.687	.290	3.449
	TIME_EXPL_reflection_2	.002	.009	.061	.171	.866	.162	6.183
	TIME_EXPL_reflection_3	-.006	.009	-.270	-.673	.509	.129	7.748
	TIME_EXPL_reflection_4	-.006	.003	-.441	-1.842	.081	.361	2.774
	TIME_management_page1	.001	.002	.090	.407	.689	.425	2.354
	TIME_UNDERST_first_story_question	.002	.002	.307	1.103	.284	.267	3.744
	TIME_UNDERST_second_story_question	.000	.001	.100	.551	.588	.626	1.598

a. Abhängige Variable: UNDERST_R_total

Dependent variable: total number of described strategy characteristics in the embedded story questions

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	Kollinearitätsstatistik	
		Regressionskoeffizient B	Standardfehler	Beta			Toleranz	VIF
1	(Konstante)	.004	.669		.007	.995		
	EXPL_quant_edu_step1	-.003	.003	-.253	-.982	.338	.260	3.842
	EXPL_quant_health_step2	.005	.003	.350	1.568	.133	.346	2.889
	EXPL_quant_roads_step3	.009	.004	.492	2.249	.037	.362	2.765
	EXPL_qual_step_1_0_no_explanation_1_wrong_explanation_of_behav	.497	.278	.344	1.792	.089	.469	2.133
	EXPL_qual_step_2_0_no_explanation_1_wrong_explanation_of_behav	.132	.536	.090	.246	.808	.130	7.702
	EXPL_qual_step_3_0_no_explanation_1_wrong_explanation_of_behav	-.367	.554	-.268	-.662	.516	.105	9.495
	EXPL_qual_step_4_0_no_explanation_1_wrong_explanation_of_behav	.485	.268	.360	1.807	.087	.435	2.297
	TIME_EXPL_exploration_step_1	.001	.002	.138	.479	.638	.208	4.805
	TIME_EXPL_exploration_step_2	-.006	.003	-.612	-1.960	.065	.178	5.630
	TIME_EXPL_exploration_step_3	.006	.004	.385	1.533	.142	.275	3.641
	TIME_EXPL_exploration_step_4	.000	.000	.088	.522	.608	.608	1.646
	TIME_EXPL_reflection_1	.002	.002	.208	.850	.406	.290	3.449
	TIME_EXPL_reflection_2	-.002	.005	-.112	-.342	.736	.162	6.183
	TIME_EXPL_reflection_3	-.007	.005	-.522	-1.425	.170	.129	7.748
	TIME_EXPL_reflection_4	-.002	.002	-.233	-1.062	.301	.361	2.774
	TIME_management_page1	.001	.001	.232	1.149	.265	.425	2.354
	TIME_UNDERST_first_story_question	.002	.001	.548	2.153	.044	.267	3.744
	TIME_UNDERST_second_story_question	.001	.000	.240	1.445	.165	.626	1.598

a. Abhängige Variable: UNDERST_S_total

Dependent variable: total number of correct answers in the multiple-choice posttest

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten			Kollinearitätsstatistik	
		Regressionskoeffizient B	Standardfehler	Beta	T	Sig.	Toleranz	VIF
1	(Konstante)	2.461	1.022		2.408	.027		
	EXPL_quant_edu_step1	-.004	.004	-.279	-.998	.331	.236	4.237
	EXPL_quant_health_step2	.007	.005	.302	1.308	.207	.346	2.889
	EXPL_quant_roads_step3	.009	.006	.359	1.471	.158	.309	3.233
	EXPL_qual_step_1_0_no_explanation_1_wrong_explanation_of_behav	.633	.414	.301	1.529	.144	.477	2.098
	EXPL_qual_step_2_0_no_explanation_1_wrong_explanation_of_behav	1.212	.809	.568	1.497	.152	.128	7.787
	EXPL_qual_step_3_0_no_explanation_1_wrong_explanation_of_behav	-1.715	.872	-.863	-1.967	.065	.096	10.434
	EXPL_qual_step_4_0_no_explanation_1_wrong_explanation_of_behav	-.086	.453	-.044	-.190	.852	.347	2.881
	TIME_EXPL_exploration_step_1	.002	.003	.176	.518	.611	.160	6.256
	TIME_EXPL_exploration_step_2	-.013	.005	-.849	-2.634	.017	.177	5.638
	TIME_EXPL_exploration_step_3	-.009	.009	-.402	-1.066	.300	.130	7.691
	TIME_EXPL_exploration_step_4	.006	.002	.973	2.862	.010	.160	6.269
	TIME_EXPL_reflection_1	-.004	.003	-.356	-1.412	.175	.291	3.439
	TIME_EXPL_reflection_2	-.002	.008	-.089	-.256	.801	.155	6.467
	TIME_EXPL_reflection_3	.012	.009	.605	1.285	.215	.083	12.025
	TIME_EXPL_reflection_4	-.003	.003	-.274	-1.131	.273	.315	3.170
	TIME_management_page1	.003	.001	.432	1.773	.093	.310	3.221
	TIME_UNDERST_first_story_question	.003	.001	.581	2.139	.046	.251	3.992
	TIME_UNDERST_second_story_question	.000	.001	-.084	-.401	.693	.424	2.359

a. Abhängige Variable: MC_post_total_correct