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

July 24-28, 2011, Washington, DC

Providing structural transparency when exploring a model's behavior: Effects on performance and knowledge acquisition

Birgit Kopainsky[1], Stephen M. Alessi[2], Pablo Pirnay-Dummer[3]

[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

[3] Institut für Erziehungswissenschaft, Albert-Ludwigs-Universität Freiburg, Rempartstr. 11, 79098 Freiburg, Germany

Abstract

Prior exploration is an instructional strategy which has improved performance and knowledge acquisition in system-dynamics based learning environments, but only to a limited degree. This study investigates whether model transparency, showing users the internal structure of models, can extend the prior exploration strategy and improve learning even more. In an experimental study, participants in a web-based simulation learned about and managed a small developing nation. All participants were provided the prior exploration strategy but only half received prior exploration embedded in a structure-behavior diagram intended to make the underlying model's structure more transparent. Participants provided with the more transparent strategy demonstrated better knowledge acquisition of the underlying model on an objective measure (multiple-choice posttest) but no difference on a subjective measure (open-ended verbal protocols based on short essay questions). Furthermore, their performance (managing the nation) was the equivalent to those in the less transparent condition. Combined with our previous studies, the results suggest that while prior exploration is a beneficial strategy for both performance and knowledge acquisition, making the model structure transparent in this way (with structure-behavior diagrams) is more limited in its effect.

Introduction and background

The difficulties with decision making in complex dynamic systems are well documented (e.g., Brehmer, 1992; Funke, 1991; Jensen, 2005; Moxnes, 1998, 2004; Rouwette, Größler, & Vennix, 2004; Sterman, 1989a; Sterman, 1989b). In previous research with system-dynamics-based learning environments (Kopainsky, Alessi, Pedercini, & Davidsen, 2009) we have shown success with an instructional strategy we call prior exploration. This strategy seeks to improve learners' performance (success in running a simulation) and knowledge acquisition (of the simulation model and strategies for working with it) by improving both their mental models and transfer of that knowledge, and simultaneously minimizing detrimental cognitive load and learners' concern with risk. Seeking to improve upon that success, we have begun a program of research to investigate other strategies to use in conjunction with the prior exploration technique. In the current study we investigate the strategy of making a model's structure more transparent so as to facilitate prior exploration.

Over the last five years we have been developing a system-dynamics based learning environment (subsequently referred to only as the learning environment) called BLEND, the Bergen Learning Environment for National Development. BLEND (developed at the University of Bergen in Norway) is based on a version of the Millennium Institute's Threshold-21 model of national development, simplified to represent the characteristics of developing nations in sub-Saharan Africa. It's learning objectives include (1) recognizing the need to balance social, economic and environmental factors in a nation's development, (2) understanding and operating within the complex non-linear dynamic relationships of such a system, (3) thinking and planning in the long term (rather than the short term) including recognition of the role played by delays, and (4) enticing learners to pursue their own modeling activities relevant to the particular characteristics of their own nations.

Our initial learning environment (Alessi, Kopainsky, Davidsen, & Pedercini, 2008) was designed to have learners play the roles of critical national leaders (the prime minister and the ministers of finance, education, health, environment, and transportation) and work cooperatively over a long (50 year) time frame to improve the nation's economic, social, and environmental conditions. Experience with that learning environment demonstrated what many designers of learning environments have reported, that understanding and working with complex system dynamics models is very difficult for learners. As a result, we have embarked on a program of research intended to improve the learning environment.

One such problem was that learners are not only overwhelmed by the complexity of decision, but are nervous, even in a game, of making the wrong decisions and seeing their simulated nation fail. A possible solution to both those problems is to give the learners a "simulator within the simulation" which would allow them to explore the model (how the nation changes when investments in areas like health, education, and infrastructure are varied) *before* they actually make decisions in the game. In Kopainsky et al., (2009) we implemented the simulator within the simulation idea using an instruc-

tional strategy we call “prior exploration”. Learners were permitted to explore the effect of individual variables (or combinations of them) on the nation, and do so quickly, easily, and without consequences for the game’s final outcome. The prior exploration strategy did improve learners’ knowledge acquisition of the model and performance in the game, but not as much as we had hoped.

Consequently, we have begun to investigate additional strategies to improve the outcomes of the prior exploration strategy. Potential strategies include giving learners corrective feedback, giving learners assignments that promote reflective thinking, using collaborative learning activities, and promoting model transparency. In our first effort, reported here, we chose model transparency as a technique for improving the previously found benefits of prior exploration.

The strategy of increasing model transparency in learning environments was actively researched in the 1990’s and early 2000’s. Since then, there has been a tendency to *assume* that model transparency is good and should be a characteristic of most learning environments and, for that matter, most system-dynamics activities (Benedetti, Bixio, Claeys, & Vanrolleghem, 2008; Crout et al., 2008; Fleischmann & Wallace, 2009; Gore, Hooey, Foyle, & Scott-Nash, 2008; Topping, Høye, & Olesen, 2010)

The research studies of Machuca and his colleagues (Machuca, Ruiz del Castillo, Domingo, & González Zamora, 1998; Machuca, 2000; González Zamora, Machuca, & Ruiz del Castillo, 2000) provided considerable evidence that well-constructed transparent models are beneficial. Several studies by Größler and his colleagues (Größler, 1997; Größler, 1998; Größler, Maier, & Milling, 2000) provided similar evidence, although some of the results were more mixed. Those and other studies are analyzed in Alessi, (2002), which in addition to concluding that transparency is beneficial for only *some* learners and *some* learning objectives, also concluded that different methods of providing transparency (verbal explanations, videos, causal-loop diagrams, stock and flow diagrams) are differentially effective. For example, stock and flow diagrams are probably effective for learners with more system-dynamics background.

Some more recent studies have again investigated (in contrast to assumed) the benefits of transparency. The results have been mixed. Cheverst et al., (2005) provided evidence that users desire transparency, though they don’t necessarily benefit from it. Cramer et al., (2008) suggested that while transparency improved users’ meta-competence (awareness of their own competence), it may have actually interfered with improving their competence. Lee, Nelles, Billinghurst, & Kim, (2004) suggested some benefits for transparency in an authoring tool, but transparency was confounded with other design characteristics, so it was not entirely clear if the benefit was due specifically to transparency. Rouwette et al., (2004) performed a literature review (including most of the studies in the previous paragraph) in which several studies of transparency *did* show beneficial results, and one very relevant study (to our work) indicated that different methods of providing transparency (e.g., causal-loop diagrams, hierarchical-tree diagrams, block diagrams) were differentially effective. Somewhat in agreement with Rouwette et al., the dissertation by Viste, (2007) included a variety of multimedia techniques for increasing transparency, some of which were more effective than others.

Given that researchers have shown success with *some* methods of increasing transparency, and based upon our theoretical belief that a key to understanding system dynamics is an appreciation how model structure drives model behavior, we chose to embed our prior exploration strategy within structure-behavior diagrams.

In our work with learning environments we are interested in two very different learning outcomes (Kopainsky, Pirnay-Dummer, & Alessi, 2010), performance and understanding (or knowledge acquisition). Performance is how well (and perhaps how quickly, though that has not been an area we have studied) learners manage the simulation. Although the national development simulation has a number of outcome variables, our main measure of performance is per capita income adjusted for interest payments on debt. It is easy for a person managing the nation to obtain high per capita incomes if they don't worry about driving the nation into debt. It is much more difficult to grow the nation in a healthy way, increasing the citizens' per capita income while avoiding national debt.

The other learning outcome, knowledge acquisition, is the extent to which the learner has internalized the simulation model as a mental model, which they can explain and base good decisions on. We measure knowledge acquisition in two ways, one objective and one subjective. Objectively, we asked multiple-choice questions both before and after using the simulation (where "using the simulation" means both the prior exploration and managing the nation in the game). Those multiple-choice questions probed knowledge acquisition of the model (e.g., the main cause-effect relationships) and about how to manage the model to produce a healthy nation (Appendix B). Subjectively, we gave the learners embedded story problem questions (Appendices D and E) for which they could type open-ended responses of whatever length they desired.

We believe that it is essential to assess both performance and knowledge acquisition when evaluating system-dynamics-based learning environments. It is possible for learners to perform well due to luck for example, without fully understanding what they are doing. It is also possible for learners to be able to explain a model, but not be able to *apply* that knowledge to problem solving, like managing the nation. We want learners to be able to solve problems or manage systems and do so for the right reasons, because they have a good mental model (knowledge) of the system.

Given the above, our research questions for this study were as follows:

1. Will learners who receive the prior exploration strategy embedded within a more transparent (structure-behavior diagram) interface show better knowledge acquisition than learners receiving the prior exploration strategy embedded in an opaque (black-box) interface? Knowledge acquisition is measured by both an objective test and by subjective open-ended essay (story) questions.
2. Will learners who receive the prior exploration strategy embedded within a more transparent interface demonstrate better performance in the final simulation-game than learners receiving the prior exploration strategy embedded in an opaque interface? Performance is measured by the final per capita income adjusted for interest payments on debt in the simulated nation.

In the remainder of the paper we refer to the group working with the more transparent interface as the transparent group and the group working with the less transparent interface as the opaque group. To answer the research questions we performed an experimental study with 144 educational psychology students. In the next section we describe the materials and methods used for the experimental study. In the results section we analyze whether the two experimental conditions differed from each other with respect to performance and knowledge acquisition. In subsequent versions of this paper the result section will also analyze the determinants of performance and knowledge acquisition by identifying those activities in the experiment that significantly influenced performance and knowledge acquisition. As our results did not find many significant performance differences between the transparent and the opaque group, the discussion and conclusions section focuses on further developments of the current experimental design.

Materials and methods

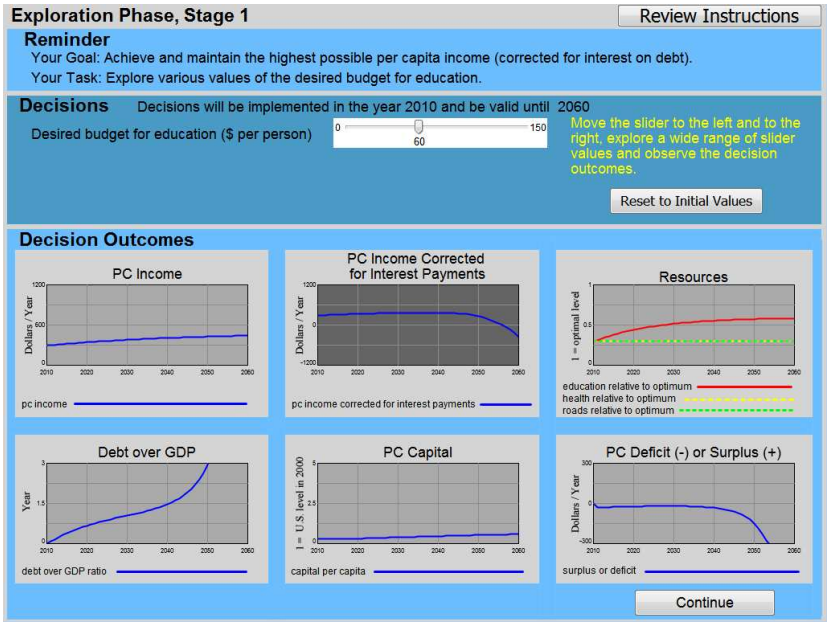
Research participants

Research participants were 144 university students from a large national university in Germany. 72 percent were female and 28% were male. 64 percent were collage age (between 18 and 21 years), 34% were 21 to 30, and 2% were above 30 years of age. Almost all were pursuing the bachelor degree. A small number (about 20 of the 144) had some experience with national development work, classes, or simulation.

Experimental conditions

The research participants were assigned randomly to one of two experimental conditions. In the original prior exploration strategy (Kopainsky et al., 2009), the learner could adjust sliders for the main input variables of the model (government expenditures for education, health, and roads) and see the effects in the form of graphs showing several of the nation's key outcome variables (e.g., national debt, per capita income, levels of education, health and roads). Figure 1 shows the original prior exploration strategy, which also served as the control condition for the study reported here (the opaque group). But in that strategy the learner only sees behavior, and nothing about the structure of the system. We therefore embedded the output graphs in a causal-loop diagram which shows the learner *both* the structure of the model *and* the behavior that results when they set the input variables (sliders) in various ways. The result, prior exploration embedded in a structure-behavior diagram, is shown in Figure 2 (the English translation) and Figure 3 (the German translation as seen by participants in Germany), which served as the experimental condition for the current study (the transparent group). The diagram also included mouse-over text. When learners point with the mouse at particular graphs, variables, arrows, or loops, they are given an explanation of their role in the overall model, intended to improve transparency even more.

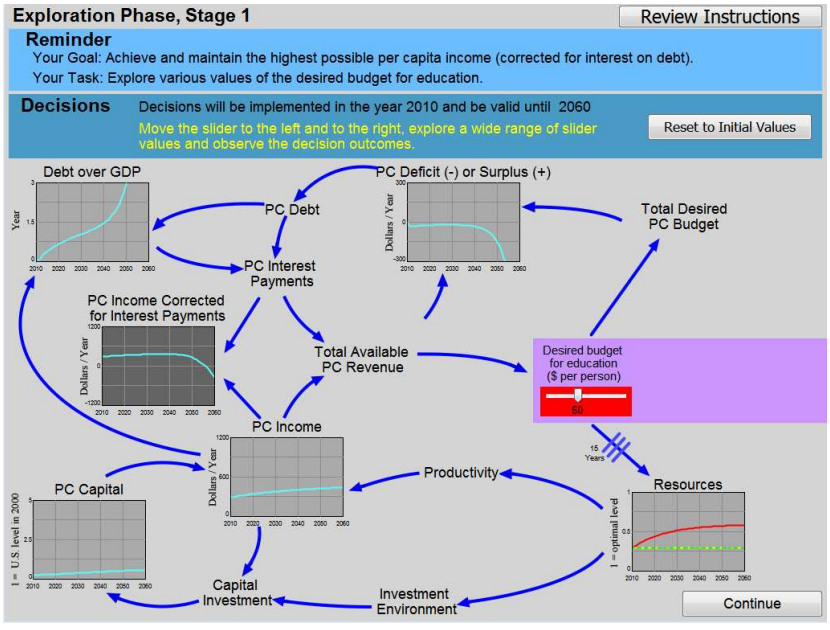
Figure 1: The Prior Exploration activity in the low transparency condition (opaque group)



Notes:

This is a dynamic activity. As the participant slides the slider for education higher and lower, the graphs below immediately replot to show how the selected budget would affect the various outcome variables. This version is considered low in transparency because there is no indication of how or why the education budget affects the variables plotted in the graphs. The exploration activity is shown in this figure in English (for the convenience of the reader). Participants in Germany saw an identical figure with the text in German.

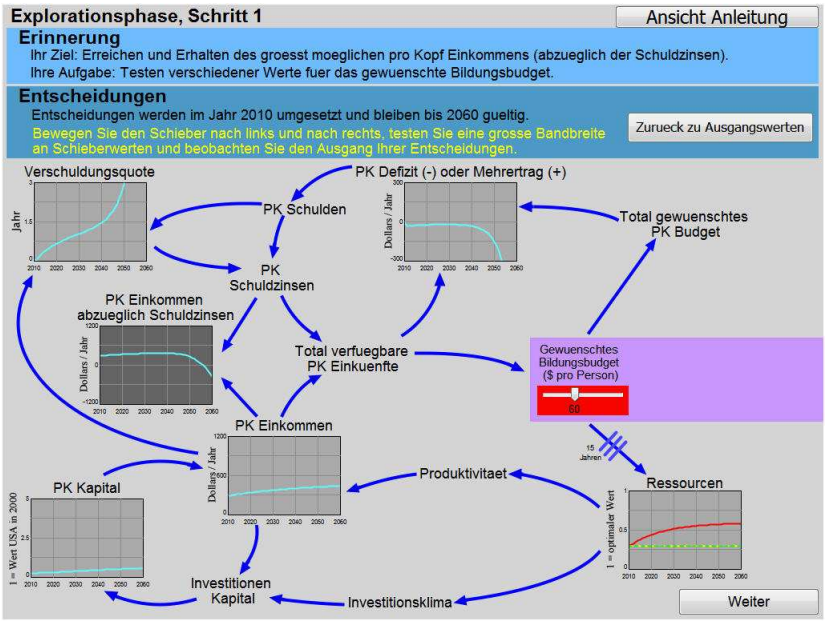
Figure 2: The Prior Exploration activity in the high transparency condition (transparency group)



Notes:

This is a dynamic activity. This version is considered high in transparency because the slider, the graphs and several other variables are shown within a causal loop diagram which reveals the cause-effect relationships. So, for example, the participant can see that the red slider for the education budget directly affects "Total Desired PC Budget" and "Resources". Those in turn affect other variables like the Deficit, Productivity, and Investment Environment. Important reinforcing loops such as the debt loop and the capital accumulation loop are also easy to see. Once again, we show an English translation, though the participants in Germany saw an identical figure with the text in German.

Figure 3: The Prior Exploration activity in the high transparency condition (transparency group) in German



Materials

All textual materials including test questions and participant responses were in German. Except for initial directions and final debriefing, all research materials were in a web-based program that could be run via any Windows-based computer with a browser and internet connection. The program consisted of

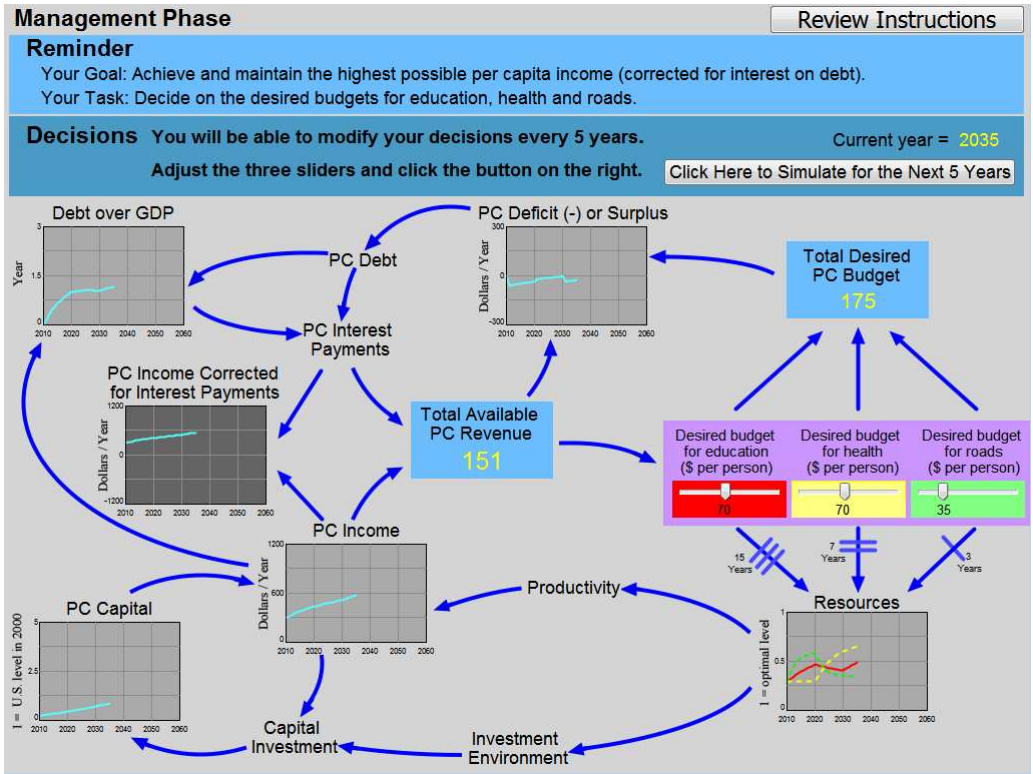
- A title page,
- Five pages of instructions (Appendix A) which described the simulated nation and the things the participants would be doing,
- An identification page which required participants to enter a unique ID number,
- An 8-item multiple-choice pretest (Appendix B),
- Four "prior exploration" stages (see below),
- The main simulation-game in which participants managed the nation for 50 simulated years (Figure 4),
- Two open-ended story questions (Appendices D and E),

- A self-assessment questionnaire (Appendix F),
- An 8-item multiple choice posttest (identical to the pretest but with the questions and their response alternatives in different order), and
- A final demographic questionnaire (Appendix G).

The four prior exploration stages were as follows.

- Participants first encountered an exploration page in which they manipulated only the expenditures for education, seeing either Figure 1 (the opaque group) or Figure 2 (the transparent group). They could do so for as long as they wanted, after which they received a reflection question as shown in Appendix C. The reflection question probed participants to type their observations about the preceding simulation-based exploration.
- Phase two was identical except that participants manipulated the expenditures for health, seeing figures very similar to either Figure 1 or Figure 2 and receiving a reflection question very similar to Appendix C.
- Phase three was the same except they manipulated the expenditures for roads (transportation infrastructure).
- Finally in phase four they were able to manipulate all three expenditure sliders for as long as they wished, once again followed by a reflection question.

Figure 4: Interface of the management phase for the transparency group



Notes:

This is the main simulation-game activity which is the basis for assessing performance. It works quite differently than the Prior Exploration activities. On this page, moving a slider does not immediately affect the graphs. 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. In this figure, the participant has so far progressed to the year 2035 (half way through the simulation), so the graphs show the nation's results up to that year. We show the English version though as with previous figures, the participants saw a version in German.

Measures

The final value of the per capita income corrected for interest payments on debt was the main measure of performance.

The pretest and posttest (Appendix B) was an objective measure of knowledge acquisition. For measurement purposes we counted the number of correct answers on the multiple-choice questions. 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). The last column refers to the levels in Bloom's taxonomy of educational objectives that are assessed with the questions. Bloom's taxonomy (Bloom, 1956; Anderson & Krathwohl, 2001) differentiates between six levels of educational objectives which start from remembering and go to understanding, applying, analyzing, evaluating, and creating. In a separate paper (Kopainsky & Alessi, submitted) we describe the taxonomy in detail and its relevance for assessing knowledge acquisition in complex dynamic decision making tasks. For the purpose of our BLEND ILE, the first four levels (remembering and understanding – levels 1/2, as well as applying and analyzing – levels 3/4) are of relevance. In the last column of Table 1 we only differentiate between levels 1/2 and 3/4, indicating questions that require remembering and explaining information about the national development planning task (levels 1/2) and questions that require using knowledge about the national development planning task to solve problems within the task (levels 3/4).

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 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	1, 2

		amount of physical capital, human capital and roads.	
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
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

The story questions (Appendices D and E) were the subjective measure of knowledge acquisition. Descriptions of the problem situation and of the proposed strategy to solve the national 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 structure of the underlying simulation model) and the strategies for successfully solving the national development planning task.

We coded a random selection of ten participants' written responses for each experimental condition and rated the responses for descriptions of relationships in the underlying simulation model and for descriptions of characteristics of successful strategies for solving the national development planning task.

As coding and rating of the verbal protocols for 144 participants would have been a very time consuming (as well as subjective) task we entered the verbal protocols into an automated analysis which we have tested for its suitability in complex dynamic decision making tasks in a previous paper (Kopainsky 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 as a result of the embedded story question. T-MITOCAR currently works with verbal protocols in either English or German.

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 stages. 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 et al., (2010).

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

Potential participants were introduced to the study during class and given the opportunity to volunteer or not for the study. Volunteers could log in for the study and, based on their student number, were randomly directed to one of two web URLs, one of which pointed to the opaque condition and the other pointed to the transparent condition of the program. Participants were allowed two weeks to perform the national development planning task. Data was automatically stored to a secure web server. After two weeks, a debriefing and discussion occurred in class.

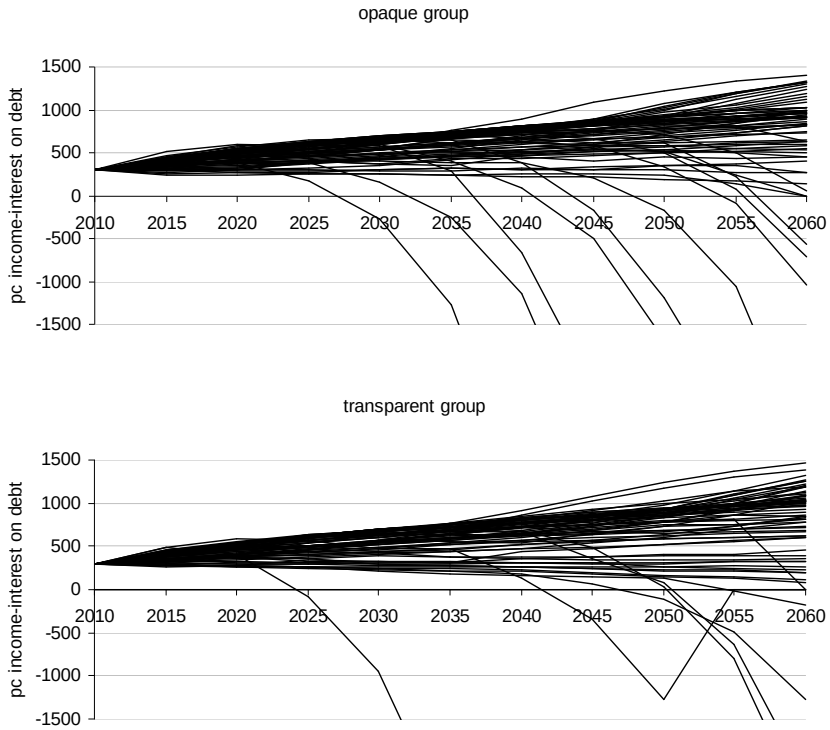
Results

This section presents the results from our experimental study. We first compare performance between the opaque and transparent group and then analyze differences in understanding knowledge acquisition by the two groups.

Performance

Figure 5 shows participants' performance in the national development planning task (i.e., the values for per capita income corrected for interest payments on debt) for the opaque group and the transparent group. In both groups, the vast majority of participants either stabilized or increased their per capita income (corrected for interest on debt) over time. About twice as many participants in the opaque group bankrupted their country, i.e., they created so much debt that per capita income corrected for interest payments on debt became negative.

Figure 5: Individual participants' performance in the two conditions



To see if the differences between the opaque group and the transparent group were statistically significant, we compared per capita income corrected for interest on debt for the two groups with two-tailed t-test at $\alpha = 0.05$. The resulting p-value for the year 2060 (the final year of the simulation) was 0.88 indicating that there was no difference in performance between the two groups based on the final per capita income corrected for interest on debt.

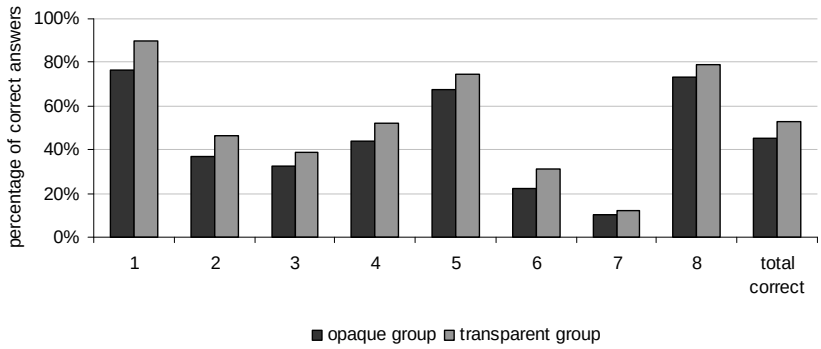
Knowledge acquisition

Figure 6 compares performance of the two conditions on the multiple-choice questions in the posttest. The figure indicates the percentage of correct answers to each question and the percentage of total correct answers for the opaque and the transparent group.

Participants performed slightly better on the pretest. However, the differences between the pre- and the posttest were not significant. Figure 6 shows that a majority of the par-

ticipants correctly answered questions about the length of the delays regarding education, health and roads expenditure (question 1), about the influence of education, health and roads on capital investment (question 5) and about the decisions in the task. Only a small number of participants were able to correctly identify the preconditions for reducing debt (question 7).

Figure 6: Multiple-choice test: percentage of correct answers in the posttest



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

Figure 6 shows that for all questions, a higher percentage of participants in the transparent group answered correctly. These differences were significant (two-tailed t-test at $\alpha = 0.05$) for question one and the total number of correct answers. Table 3 provides detailed statistics for the differences between the opaque and the transparent group in the posttest. In addition to the percentage of total correct answers the table also lists the percentage of correct answers to level 1/2 questions and to level 3/4 questions. When the questions are split into level 1/2 and 3/4, the differences between the opaque and the transparent group are not significant anymore (at $\alpha = 0.05$). However, the table shows that there is a tendency for the transparent group to outperform the opaque group for the higher level questions (level 3/4).

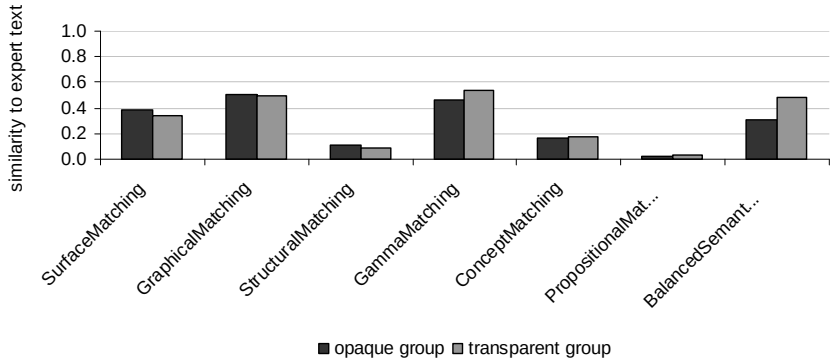
Table 3: Results from a two-tailed t-test concerning differences between the opaque and transparent group in the multiple-choice posttest

	average % of total correct answers	average % of correct answers level 1/2 questions	average % of correct answers level 3/4 questions
opaque group	45	50	41
transparent group	53	58	48
p values	.04	.12	.07

Figure 7 presents the results from the automated analysis of the verbal protocols. 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 equal to the expert response for a specific structural or semantic characteristic.

Figure 7: Structural and semantic similarity between the verbal protocols and the 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 describe a fair number of concepts (variables) in their responses (fairly high level for surface matching), and that they link these concepts quite intensively (high levels for graphical matching and gamma matching). The low values for concept and propositional matching, however, indicate that the concepts that they describe are not very important in the national development planning task (i.e., they show a low level of concept matching to the expert response) and that they do not link the concepts correctly (i.e., they show a low level of propositional matching to the expert response).

The opaque and the transparent group differ from each other significantly for the gamma matching index (two-tailed t-test at $\alpha = 0.05$). The transparent group thus showed a level of interconnectedness of concepts (variables) that was closer to the expert response than that of the opaque group.

For assessing participants' knowledge acquisition we also coded some of the verbal protocols manually. Manual analysis was only performed for ten protocols per experimental condition. The manual analysis identified the number of described relationships in the verbal protocols and the number of described strategy elements for solving the national development planning task. The two experimental conditions did not differ from each other significantly, neither in terms of relationships nor strategy elements (based on a Mann-Whitney t-test at $\alpha = 0.05$). It is, however, possible that the lack of significant differences is entirely caused by the low number of analyzed protocols. In subsequent versions of this paper we will increase the number of manually coded protocols per condition to increase the statistical power for this assessment.

Discussion and reflection

Research questions

Our *first research question* was the following. Will learners who receive the prior exploration strategy embedded within a more transparent interface show better knowledge acquisition than learners receiving the prior exploration strategy embedded in an opaque interface?

We employed two measures to address this question. The first, an objective measure, was an eight-item multiple-choice test given before and after the simulation activities (exploration and management of the model). Four of the items probed participants' knowledge acquisition at the first and second levels of Bloom's taxonomy of educational objectives (Bloom, 1956; Anderson & Krathwohl, 2001) (remembering and understanding). The other four items probed participants' knowledge acquisition at the third and fourth levels of Bloom's taxonomy (applying and analyzing).

The second, a subjective measure, consisted of two short-essay story questions given immediately after the final simulation activity (management of the model). The first story question asked participants to write a note to the prime minister describing the problem facing the nation, that is, explaining the main issues and variables relevant to the nation and how they affect each other. The second story question followed up on the first, asking participants to advise the prime minister by suggesting an investment strategy (for education, health, and roads across a 50-year time span) to maximize per capita income while minimizing national debt.

The objective multiple-choice test given before the simulation activities (the pretest) showed no significant difference between the participants given a more transparent model interface and those given a more opaque model interface. The objective multiple-choice test given after the simulation activities (the posttest) did demonstrate a significant ($p=.04$) difference favoring participants receiving the more transparent model interface. Those in the transparent condition answered an average of 53% of the questions correctly in the posttest, while those in the opaque condition answered an average of 45 percent correctly.

Somewhat surprisingly, overall performance on the pretest was marginally *better* than on the posttest, but that difference was not significant. Nor was there any significant difference on the pretest between conditions or for different types of questions. Because the pretest-posttest difference was not significant, we would not conclude that participants did worse on the posttest, however, we also cannot conclude that they did better. We can only say that after the simulation activities, those in the transparent condition performed better. Let's consider why this might be the case.

The pretest was given before the simulation activities, but *after* the instructions. Those instructions included a description of the nation of Blendia, information about key variables (investments in education, health, roads), and issues like revenue, borrowing, and interest payments on debt. In other words, some instruction *was provided* in the instruc-

tions, though only in the form of participants reading verbal information. Only later during the simulation activities did they work with the information learned. The pretest probably reflects *some* learning that occurred from reading the instructions. The posttest, in contrast, reflects the more significant learning that occurred from *both* the instructions (reading about the variables and issues) and the simulation activities (actually experimenting with and manipulating the variables). We had included a pretest in the hope that it would provide greater statistical power by taking entry knowledge into consideration, but it did not appear to do that since the two conditions did not differ at all on the pretest. The posttest turned out to be the best indication of overall learning from both instructional and simulation activities.

Because the posttest items were of two types, four at the remembering and understanding levels of Bloom's taxonomy (levels 1 and 2) and four at the slightly higher applying and analyzing (3 and 4) levels, we also examined how the two conditions differed for the different levels of questions. With only four (instead of eight) questions the significant differences disappeared, but the trend continued and was greater for the level 3 and 4 questions. That is, while participants in the transparent condition did only marginally better on level 1 and 2 question than did participants in the opaque condition ($p=.12$), for the level 3 and 4 questions the transparent condition showed greater improvement over the opaque condition, with $p=.07$ being almost significant. A more challenging test, perhaps with more questions at the application and analysis levels, might have demonstrated a significant difference. Our cautious (given that these differences were not significant) new hypothesis is that model structure transparency benefits higher levels of learning (applying and analyzing) more than lower levels of learning.

Unfortunately, that hypothesis is only marginally supported by our subjective (story problem) questions. The automated analysis of the participants' verbal protocols on the story problems (done by the T-MITOCAR program) showed only one significant difference between the transparent and opaque conditions and the manual analysis showed no significant differences. T-MITOCAR calculates seven different indices of similarity (between the participants' answers and experts' answers to the same questions), and the only index that was significantly different was Gamma Matching. This reflects the amount of interconnectedness among concepts in a response, and indicated that participants in the transparent condition had interconnectedness more like that of experts than did participants in the opaque condition.

Why would the objective multiple-choice posttest show more differences than the subjective story questions? The most obvious answer is that the objective questions are more focused on key concepts and therefore more sensitive to differences in knowledge acquisition concerning those concepts. Participants' responses to essay questions are all over the place, often not addressing the key concepts at all. The much greater variation makes detecting differences among conditions more difficult.

Our *second research question* was the following. Will learners who receive the prior exploration strategy embedded within a more transparent interface demonstrate better performance in the final simulation-game than learners receiving the prior exploration

strategy embedded in an opaque interface, where performance is measured by the final per capita income adjusted for interest on debt in the simulated nation.

Using the criterion of per capita income minus interest on debt in 2060 (the last year of the management simulation), the two conditions did not differ significantly. Looking at Figure 5, we see that the great majority of participants in both groups had small to large improvements based on that criterion. A small number of participants did poorly, bankrupting the nation, as represented by lines going down below zero in the two graphs. In fact, adopting a "bankruptcy" criterion (whether or not participants bankrupt the nation), the transparent condition appear to perform better. Only about five participants in that condition bankrupted the nation. In contrast, nine participants in the opaque condition bankrupted the nation, almost twice as many. However, these numbers are too small to demonstrate a significant difference. They only suggest that while the great majority of participants perform well, transparency may reduce the small number of very poor performances (bankruptcies).

Reflections

Given some success regarding our first research question but much weaker findings regarding the second research question, our main question is why might learners acquire relevant knowledge yet not perform well within the simulation? The most obvious answer is that performance requires transfer of knowledge from one form (answering verbal questions) to another (policy formation and implementation). It is quite common for learners to acquire new knowledge yet not be able to apply it in other situations, especially in the real world. It makes sense that providing learners with transparent model structure, including showing how that structure relates to model behavior (in the form of the output graphs), would help them understand the model better. In fact, they appear to understand the model not only at the *simplest* levels of Bloom's taxonomy (remembering and understanding) but even more so at the slightly *higher* levels of applying and analyzing. But *applying* in a multiple-choice question is not the same as applying when implementing policies and strategies in a management simulation over a period of fifty (simulated) years. No matter what the level of knowledge required in a multiple-choice question, the learner still need only click on a response. To be successful in the management simulation probably requires learners to form hypotheses, test them, evaluate the results, and revise hypotheses, doing all that several times. We know from our previous experiments that the prior exploration strategy does itself impact performance, but simply modifying its interface (providing greater or less transparency) mostly impacts knowledge acquisition, and impacts performance little, if at all.

Given the overall performance of our participants (some still bankrupt the simulated nation and many just hold the nation steady, without improving anything) we are certain that they can still improve a lot. Research on model structure transparency suggests it is sometimes beneficial. But there are other ways to provide structural transparency besides imbedding behavior graphs in a causal loop diagram. The structure of a model could be taught with an interactive tutorial, with an audio or video lecture, with animated pedagogical agents, or any number of new multimedia techniques. Our structure-

behavior diagrams were passive, that is, learners were not *required* to cognitively process the information embedded in them. Perhaps a form of structural transparency which requires more active cognitive processing will be more effective.

Then again, perhaps the prior exploration strategy will be augmented more by something other than transparency of model structure. For example, providing assistance (either through a help system or an animated pedagogical agent) on *exploring* (creating hypotheses, testing them, revising them) might have even greater impact than providing structural transparency tools.

Next steps

Although not all our hypotheses were confirmed and not all our measures were effective, results were sufficient to suggest modifications to our research with the current learning environment. The pretests did not add much information, so can probably be eliminated. The story questions might be asked immediately after the exploration phases, which would provide a more sensitive test of how exploration affects knowledge acquisition. Given the procedure we used, the posttest and story questions followed *both* the exploration and the management, so the effect of exploration (with or without transparency) may have been diluted by additional learning during the management phase. Finally, simply providing information about model structure (transparency) does not guarantee that learners cognitively process it, so we plan to investigate more interactive strategies which encourage greater processing of the structure-behavior diagrams.

References

Alessi, S. M. (2002). *Model transparency in educational system dynamics*. Paper presented at the 20th International Conference of the System Dynamics Society.

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.

Benedetti, L., Bixio, D., Claeys, F., & Vanrolleghem, P. A. (2008). Tools to support a model-based methodology for emission/immission and benefit/cost/risk analysis of wastewater systems that considers uncertainty. *Environmental Modelling & Software, 23*(8), 1082-1091.

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

Brehmer, B. (1992). Dynamic decision making: Human control of complex systems. *Acta Psychologica, 81,* 211-241.

Cheverst, K., Byun, H. E., Fitton, D., Sas, C., Kray, C., & Villar, N. (2005). Exploring issues of user model transparency and proactive behavior in an office environment control system. *User Modeling & User-Adapted Interaction, 15,* 235-273.

Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., et al. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling & User-Adapted Interaction, 18*(5), 455-496.

Crout, N., Kokkonen, T., Jakeman, A. J., Norton, J. P., Newham, L. T. H., Anderson, R., et al. (2008). Good Modelling Practice. In A. J. Jakeman, A. A. Voinov, A. E. Rizzoli & S. H. Chen (Eds.), *Developments in Integrated Environmental Assessment*: Elsevier.

Fleischmann, K. R., & Wallace, W. A. (2009). Ensuring transparency in computational modeling. *Communications of the ACM, 52*(3), 131-134.

Funke, J. (1991). Solving complex problems: Exploration and control of complex systems. In R. Sternberg & P. Frensch (Eds.), *Complex problem solving: principles and mechanisms* (pp. 185-222). Hillsdale, NJ: Lawrence Erlbaum.

González Zamora, M. D. M., Machuca, J. A. D., & Ruiz del Castillo, J. C. (2000). SITMECOM 1.0 PC: A transparent-box multifunctional simulator of competing companies. *Simulation & Gaming, 31*(2), 240-256.

Gore, B. F., Hooey, B. L., Foyle, D. C., & Scott-Nash, S. (2008). *Meeting the challenge of cognitive human performance model interpretability through transparency: MIDAS v5.x*. Paper presented at the 2nd International Conference on Applied Human Factors and Ergonomics.

Größler, A. (1997). *Giving the black box a lid – Providing transparency in management simulations*. Paper presented at the 15th International Conference of the System Dynamics Society.

Größler, A. (1998). *Structural transparency as an element of business simulators*. Paper presented at the 16th International Conference of the System Dynamics Society.

Größler, A., Maier, F. H., & Milling, P. M. (2000). Enhancing learning capabilities by providing transparency in business simulators. *Simulation & Gaming, 31*(2), 257-278.

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

Jensen, E. (2005). Learning and transfer from a simple dynamic system. *Scandinavian Journal of Psychology, 46*(2), 119-131.

Kopainsky, B., & Alessi, S. M. (submitted). *Measuring knowledge acquisition in dynamic decision making tasks*. Paper presented at the 29th International Conference of the System Dynamics Society.

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., 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.

Lee, G. A., Nelles, C., Billinghurst, M., & Kim, G. J. (2004). *Immersive authoring of tangible augmented reality applications.* Paper presented at the International Symposium On Mixed And Augmented Reality (ISMAR).

Machuca, J. A. D. (2000). Transparent-box business simulators: An aid to manage the complexity of organizations. *Simulation & Gaming, 31*(2), 230-239.

Machuca, J. A. D., Ruiz del Castillo, J. C., Domingo, M. A., & González Zamora, M. D. M. (1998). *Our ten years of work on transparent box business simulation.*

Paper presented at the 16th International Conference of the System Dynamics Society.
Moxnes, E. (1998). Not only the tragedy of the commons: Misperceptions of bio-economics. *Management Science, 44*(9), 1234-1248.
Moxnes, E. (2004). Misperceptions of basic dynamics: the case of renewable resource management. *System Dynamics Review, 20*(2), 139-162.
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.
Rouwette, E. A. J. A., Größler, A., & Vennix, J. A. M. (2004). Exploring influencing factors on rationality: a literature review of dynamic decision-making studies in system dynamics. *Systems Research and Behavioral Science, 21*(4), 351-370.
Sterman, J. D. (1989a). Misperceptions of feedback in dynamic decision making. *Organizational Behavior and Human Decision Processes, 43*(3), 301-335.
Sterman, J. D. (1989b). Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Science, 35*(3), 321-339.
Topping, C. J., Høye, T. T., & Olesen, C. R. (2010). Opening the black box – Developing, testing and documentation of a mechanistically rich agent-based model. *Ecological Modeling, 221*, 245-255.
Viste, M. (2007). *Visualization of Complex Systems.* University of Bergen, Bergen, Norway.

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: Exploration workbook - Part 1

What happened to per capita income and the other indicators when you changed the budget for education?

Why do you think this happened?

Please write your key observations below.

Appendix D: Embedded story question - Part 1

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 E: Embedded story question - Part 2

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 F: Briefing the Prime Minister - Part 3

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

	I strongly	I disagree	I neither dis-	I agree	I strongly

	disagree		agree nor agree		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 G: 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