Paper presented at the 28th International Conference of the System Dynamics Society

July 25-29, 2010, Seoul, Korea

AUTOMATED ASSESSMENT OF LEARNERS' UNDERSTANDING IN COMPLEX DYNAMIC SYSTEMS

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

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

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

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

Abstract

Research on learning via system-dynamics-based learning environments depends on accurate measurement of learning. Most such research considers at least two aspects of learning, the participants' understanding of the models and problems, and the participants' performance in the environment, e.g., quality of decision making. The former, understanding, is much more difficult to measure than the latter, performance. Measurement of understanding is often done by eliciting verbal protocols from participants about the problem situation (i.e., the underlying model) and their planned solution strategy (i.e., decisions). Coding and analysis of participants' verbal protocols is very subjective and time-consuming. To facilitate measurement and analysis of understanding via verbal protocols, we investigate the utility of a software application which performs such analysis automatically. We assess this automated analysis methodology using data from two different system-dynamics-based learning environments and analyze how participants' understanding compares to experts, how it changes over time, and how it correlates with performance.

1 Introduction

Dynamic systems such as the economy of a country, production in companies, renewable resources or global warming are difficult to understand and manage successfully (Diehl & Sterman, 1995; Jensen & Brehmer, 2003; Jensen, 2005; Moxnes, 1998; Moxnes, 2004; Paich & Sterman, 1993; Sterman, 1989; Sterman, 2002 & 2007). One of the primary goals of system dynamics is to improve decision making and problem solving in complex dynamic systems. In order to know when we have improved decision making or problem solving, we must have valid and reliable methods to assess how well people do them. Such assessments usually focus on two measures of the people involved (Rouwette, Größler, & Vennix, 2004): their performance, i.e., the results from decision making (such as, a score on a gaming variable); and their understanding, i.e., the rules and mental operations that lead people to their decisions. Especially when a strategy seems to yield promising results in terms of performance, it becomes essential to know whether improved performance is due to the person's improved understanding of the system or due to other reasons such as trial and error.

Our current work and this paper is part of the emerging literature in system dynamics that seeks to assess system understanding or elicit people's mental models, respectively (Capelo & Dias, 2009; S. Cavaleri & Sterman, 1997, Doyle, 1997; Jensen & Brehmer, 2003; Jensen, 2005; Huz, Andersen, Richardson, & Boothroyd, 1997; Spector, Christensen, Sioutine, & McCormack, 2001).

A mental model is a representation of a thing, ideas or more generally an ideational framework. Representations are widely viewed as having a language-like syntax and a compositional semantic (see Carruthers, 2000; Fodor, 2003; Margolis & Laurence, 1999; Pinker, 1994; Strasser, in press). Mental models, as types of representations, rely on language and use symbolic pieces and processes of knowledge to construct a heuristic for a situation (see Johnson-Laird, 1983; Schnotz, 1994; Schnotz & Preuss, 1997; Seel, 1991). Their purpose is heuristic reasoning which leads to either intention, planning, behavior or to a reconstruction of cognitive processes (see Piaget, 1976).

People construct mental models to match the behavior of both predictable and unpredictable changes in the world in order to exercise better control and make the changes more predictable. This also is a key aspect of much problem solving, including the complex problem solving typical of most dynamic decision making tasks (Ceci & Ruiz, 1992; Jonassen, 2000; Just & Carpenter, 1976; Spector, 2006).

A re-representation is an external correlate to a representation. It may be constructed to support learning (see Hanke, 2006) and for assessment (see Pirnay-Dummer, 2006; Pirnay-Dummer, Ifenthaler, & Spector, 2010; Ifenthaler, 2006; Ifenthaler, Masduki, & Seel, 2009; Johnson, O'Connor, Spector, Ifenthaler, & Pirnay-Dummer, 2006; Johnson et al., 2009). We call those constructs re-representations to illustrate that they are representations of representations. For example, cognitive models (such as a computer animation of a physical process) are re-representations of mental models (a person's mental representation of the same physical process).

Verbal protocols play an important role in the process of exploring people's reasoning processes (e.g., Sterman, 2009). Verbal protocols are particularly suitable for characterizing mental models as they closely approximate the way people naturally go about representing their knowledge (Doyle, Radzicki, & Trees, 2008). Unfortunately, coding and rating verbal protocols for the purpose of representing mental models is extremely challenging as the persons doing the rating require a good understanding of the phenomenon under study. This makes it difficult to find more than one suitable rater (Jensen, 2005), especially when the phenomenon is one depicted by a complex system-dynamics model. Furthermore, coding and rating of verbal protocols are subject to interpretation by the raters (e.g., Sterman & Booth Sweeney, 2007) and are time consuming tasks. In short, human coding and rating of verbal protocols is very difficult for large datasets, such as when many research participants produce them as a measure of understanding in a decision-making or problem-solving activity. Thus, automated or semi-automated coding and analysis of verbal protocols would be a valuable research tool for assessing people's understanding in dynamic decision making tasks.

In this paper we investigate such an automated analysis of textual understanding data (verbal protocols) and whether it can improve assessment in complex dynamic systems. The particular automated analysis of textual understanding data is based on T-MITOCAR (Pirnay-Dummer & Spector, 2008; Pirnay-Dummer & Ifenthaler, 2010), which stands for Text – Model Inspection Trace of Concepts and Relations. The software is based on mental model theory (Seel, 1991) and uses syntactic and semantic heuristics to track associations of terms within written language. As will be discussed more in the methods section, there are several advantages implicit in the design of T-MITOCAR. The texts that are used as input need not be coded prior to analysis. Also, T-MITOCAR does not require any domain dependent linguistic or structural corpus to analyze the texts. T-MITOCAR has shown to be stable across domains and in many different settings of learning (see Pirnay-Dummer, 2006, 2007, 2008; Pirnay-Dummer & Ifenthaler, 2010). Moreover, first studies indicate that the T-MITOCAR models can also be used to predict learning progression over time with surprisingly high correlations (e.g., r=.99, see Schlomske & Pirnay-Dummer, 2009). Also, considerable correlations where found when the methodology was compared to other language oriented assessment strategies like Latent Semantic Analysis (see Pirnay-Dummer & Walter, 2009).

The main purpose of the work reported here is therefore to test whether the automated analysis is also valid for assessing understanding of complex dynamic problems. The focus of understanding assessment in the system dynamics literature is mainly on the evaluation of systems thinking skills (S. A. Cavaleri & Thompson, 1996; Hopper & Stave, 2008; Maani & Maharaj, 2001; Richmond, 1997; Skaza & Stave, 2009, Skaza & Stave, 2010). In this paper we concentrate on evaluating the ability of learners to appropriately describe the problem situation and to explain a solution strategy for a dynamic decision making task. The interpretation of such descriptions applies systems thinking characteristics. The first step, thus, is to automate the analysis of verbal protocols and test how valid such analysis is for complex dynamic problems. Ultimately, a second step would then further develop the automated analysis such that it can interpret a text in systems thinking terms. In other words, the goal in the long run would be to correlate task specific descriptions with the specific systems thinking characteristics represented by them.

To investigate the validity of the automated analysis we use textual understanding data from two experimental dynamic decision making tasks and compare the results of the automated analysis to results generated by a manual analysis of understanding data. We also analyze how participants' understanding compares to experts, how it changes over time, and how it correlates with performance. We conclude with a theoretical and methodological discussion of how an automated analysis of verbal protocols can help to identify more clearly the misperceptions that lead to suboptimal decisions.

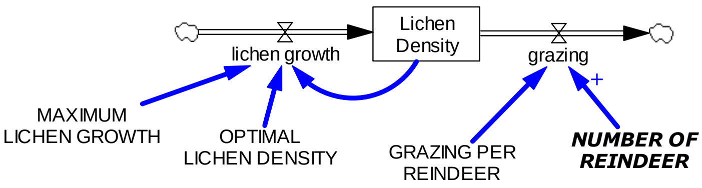
2 Dynamic decision making tasks

The first task, the reindeer rangeland management task (Moxnes, 2004) is based on a relatively simple one-stock model. The system dynamics model underlying the second task, the national development planning task (Kopainsky, Alessi, Pedercini, & Davidsen, 2009), contains five stocks and is thus more complex. It should be emphasized, however, that even though the first task has only one stock and is comparatively quite simple, learners usually fail to solve the task, even learners with experience in system-dynamics modeling or natural resource management (Moxnes, 1998, Moxnes, 2004).

2.1 Reindeer rangeland management task

The reindeer rangeland management task was developed by Moxnes, (2004). In this task, participants play the role of sole owners of a reindeer herd. They take over the herd and overgrazed rangeland from a previous owner, and are responsible for setting the reindeer herd size for each of 15 simulated years. Participants' goal is to restore the maximum sustainable herd size as quickly as possible. The instructions they receive (cf. appendix 1) provide information about the grazing rate of the reindeer and a description of lichen growth dynamics. Lichen is the source of food to support reindeer through the winter and is therefore a limiting factor for the size of a herd. The instructions contain a 15-year long historical record on lichen density and reindeer herd sizes. The simulation model underlying the task contains one stock (lichen density) which increases with lichen growth and decreases with grazing (Figure 1). Lichen growth is a convex and thus non linear function of lichen density.

Figure 1: Stock-flow model of the reindeer rangeland management task



Data for the reindeer rangeland management task was collected with 129 environmental science undergraduate students at the University of Nevada in Las Vegas in the fall 2009. Students studied the instructions shown in appendix 1 and were asked to explain the problem they were faced with in this task and to propose a strategy to solve the task. They then had three decision making trials (attempts) in which they implemented their strategy in a simulation. After the first trial they were given the opportunity to modify their explanations. They were given the same opportunity once again after completing all three decision making trials. The resulting dataset used for the analysis in this paper therefore consists of the participants' textual descriptions (verbal protocols) for three measurement time points (after instructions but before interacting with the simulation game, after trial 1, after interacting with the simulation game) and the participants' performance data for decision making trials 1, 2, and 3. Performance was measured by subtracting actual lichen density from optimal lichen density for the decision making period of 15 years. In other words, the closer the lichen density was to the optimal density, the better their performance.

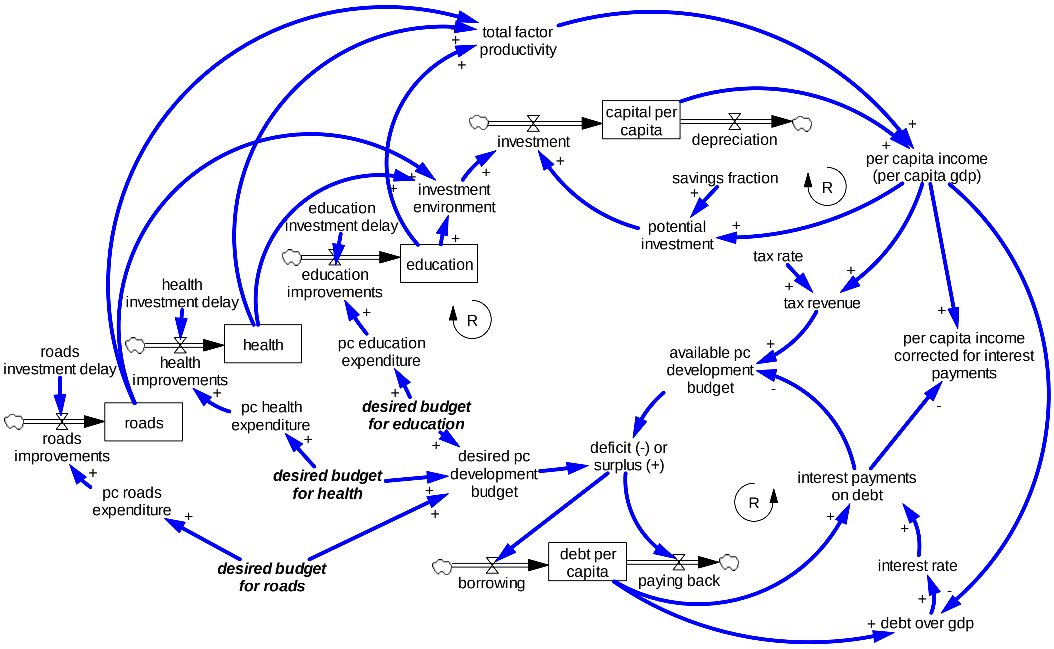
2.2 National development planning task

In the national development planning task (Kopainsky, et al., 2009) participants play the role of 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 is to achieve and maintain the highest possible per capita income over the relatively long period of 50 years (cf. appendix 2 for the complete instructions). The prime minister has far reaching financial responsibilities and the absolute power to make the following decisions:

- Investment in education (an explicit decision)
- Investment in health (an explicit decision)
- Investment in roads (an explicit decision)
- Borrowing to finance such investments (an implicit decision resulting from the three previous ones and the nation's available budget).

The simulation model used for the task (Figure 2) depicts the development of per capita income over time as a consequence of reinforcing economic growth processes between capital accumulation resulting from private sector development, infrastructure (roads) and human development (education and health). The money available for investments in roads, education and health is generated through taxes and borrowing. Borrowing creates a deficit which accumulates debt over time. Interest payments on debt are deducted from tax revenue so that the reinforcing debt accumulation loop can counteract the reinforcing economic growth loops.

Figure 2: Stock-flow model of the national development planning task



Data for the national development planning task was collected with 34 introductory level system dynamics students in the fall semester of 2009. Students studied the instructions shown in appendix 2 and were asked to explain the problem they were faced with in this task and to explain the strategy they planned to use to solve the task. They had two decision making trials in which they implemented their strategy using a system-dynamics based simulation game. After the second trial they were given the opportunity to modify their explanations. The resulting dataset used for the analysis in this paper therefore consists of the participants' textual descriptions for two measurement time points (before and after interacting with the simulation game) and of performance data for decision making trials 1 and 2. Performance was measured by subtracting per capita interest payments from per capita income for the decision making period of 50 years. In other words, better performance was represented by higher per-capita income *not* due to borrowing, but due to true economic growth.

3 Method

The analysis of a participant's textual description involves a comparison of the participant's text with an expert text. This holds true both for the automated and the by-hand analysis. Experts categorize problems more precisely, managing to find the important aspects that are relevant to a situation. Klein's (1997) recognition primed decision making model posits that experts do not chose among alternatives, but rather assess the nature of the situation and, based on this assessment, select an action appropriate to it. The first step in Klein's recognition model is to classify the situation as typical or novel. To recognize the situation, the decision maker identifies critical cues that mark the type of situation and causal factors that explain what is happening and what is going to happen. Based on this, the expert sets plausible goals and proceeds to selecting an appropriate course of action.

An important first step in the analysis and comparison of participants' mental models is therefore to elicit experts' understanding of the dynamic decision making tasks used in this paper. Eliciting expert understanding is closely related to the idea analysis applied in Jensen & Sawicka, 2006; and Booth Sweeney & Sterman, 2000, to the task analysis step applied e.g., in Jensen & Brehmer, 2003 and Jensen, 2005, and to eliciting expert conceptualizations of the problem space (Spector, 2006). For both decision making tasks a panel of experts wrote their own descriptions of the problem situation and the strategy to solve the problem. The expert texts are listed in appendix 3.

3.1 Automated analysis with T-MITOCAR

T-MITOCAR is a software tool that uses natural language expressions (instead of graphical drawings by subjects) 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 texts written by research participants (subjects) as a result of some writing task, e.g., the task of describing the problem situation underlying a decision making task and the proposed strategy to solve the decision making task.

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 a heuristic to do so. Closer relations tend to be presented more closely within a text. This does not necessarily work within single sentences, since syntax is more expressive and complex. But 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. The two basic features of T-MITOCAR are:

- Re-representation of mental models through association nets of concepts and relationships found in and generated from a written text. The association net is the result of the re-representation process described in detail in section 3.1.1.
- Analysis and comparison of mental models in terms of their structural and semantic characteristics. A subject's text may be compared to any expert text, teacher text, or any standard or model solutions for a task. It is also possible to track change over time (if subjects write texts at several measurement time points) or to measure semantic and structural differences within or between groups or subjects. The structural, semantic and combined indices used for the comparison of mental models are described in section 3.1.2.

3.1.1 Re-representation process and association nets

The re-representation of a mental model in the form of an association net is carried out in different stages (Table 1). All of the stages are automated. Thus, the only data needed is a text written by a subject such as an expert, a learner, or a teacher. We illustrate the re-representation process with experts' textual descriptions of the optimal strategies for solving the two dynamic decision making tasks introduced in section 2.

Table 1: Re-representation process in T-MITOCAR

1	Preparing the text	When text enters the system from sources unknown to the software, it most often contains characters which could disturb the re-representation process. Thus, a specific character set is expected. All other characters and formatting code are deleted.
2	Tokenizing	After preparation, the text is split into sentences and tokens. Tokens are words, punctuation marks, quotation marks, and so on.
3	Tagging	Only nouns and names should be part of the final re-representation graph. Tagging helps to find out which words are nouns or names.
4	Stemming	Different inflexions of a word appear only once in the re-representation graph. Stemming reduces all words to their word stems. All words in the initial text and all words in the tagged list of nouns and names are stemmed before the re-representation.
5	Retrieving the most frequent concepts	After tagging and stemming, the most frequent noun stems (concepts) are listed from the text.
6	Calculating the degree of association between concepts	The degree of association between concepts is calculated in several steps: 1. Calculation of the default length. For each sentence the words are counted. The default length is the number of words in the longest sentence within the text plus 1. 2. All retrieved concepts are paired, so that all possible pairs of concepts are in a list. 3. For each pair all sentences are investigated. If the pair appears within a sentence, the distance for the pair is the minimum number of words between the terms of the pair within the sentence. If at least one term occurs more than one time in the sentence, then the lowest possible distance is taken. 4. If a pair does not appear in a sentence (true also if only one concept of the pair is in the text), then the distance will be the default length. 5. The sum of distances is determined for each pair. 6. The N pairs with the lowest sum of distances are included in the re-representation graph. N depends on the number of words and sentences within the text. The exact values can be controlled by the software settings. 7. The algorithm automatically truncates the maximum distance for re-representation. This prevents the algorithm from generating random pairs which do not really have any association evidence within the text.
7	Determining the weights	After determining the degree of association between pairs of concepts, the weights are calculated from the pair distances. All weights ($0 \leq w \leq 1$) are linearly mapped so that 1 is the pair with the lowest sum of distances and 0 is the pair with the maximum sum of distances.
8	De-stemming	From the list of words and their stems produced in step 4 T-MITOCAR searches for the inflection of the word that most frequently led to the stem: If it was the plural then the plural is presented in the re-representation graph.
9	List form (for an example see Table 2)	T-MITOCAR constructs a table with the pairs of concepts, their sum of distances, and their weights (association strength).
10	Association net (for an example see Figure 3)	The graphical output of the re-representation process is the association net that displays the most important concepts and their association strengths on the basis of the list form.

Based on an expert’s verbal description of the problem situation and of the optimal strategy to solve the national development planning task, i.e., the description of the structure of the simulation model (see Figure 2) and some of the behavior this structure gives rise to (detailed expert text in appendix 3), T-MITOCAR generates the list form presented in Table 2.

Table 2: List form of the national development planning optimal strategy description

Concept 1	Concept 2	Distances	Weight
education	health	57	1
roads	education	60	0.909091
roads	health	60	0.909091
capita	debt	66	0.727273
capita	income	72	0.545455
roads	investment	75	0.454545
capita	interest	75	0.454545
education	investment	75	0.454545
capita	development	75	0.454545
health	investment	75	0.454545
capita	budget	75	0.454545
debt	interest	75	0.454545
budget	development	75	0.454545
interest	payments	78	0.363636
capita	payments	78	0.363636
debt	payments	78	0.363636
investment	environment	78	0.363636
capita	deficit	81	0.272727
tax	revenue	81	0.272727
capita	borrowing	81	0.272727
roads	environment	81	0.272727
roads	levels	81	0.272727
roads	years	81	0.272727
education	environment	81	0.272727
education	levels	81	0.272727
health	environment	81	0.272727
health	levels	81	0.272727
capita	roads	81	0.272727

The pair distance values in Table 2 have no direct meaning. They depend on the longest sentence of the text. Therefore, only the relative measure of the weights is directly interpretable as it represents the strength of association between two concepts. The list form is transformed into the association net (re-representation graph) shown in Figure 3.

Figure 3: Association net of the national development planning optimal strategy description

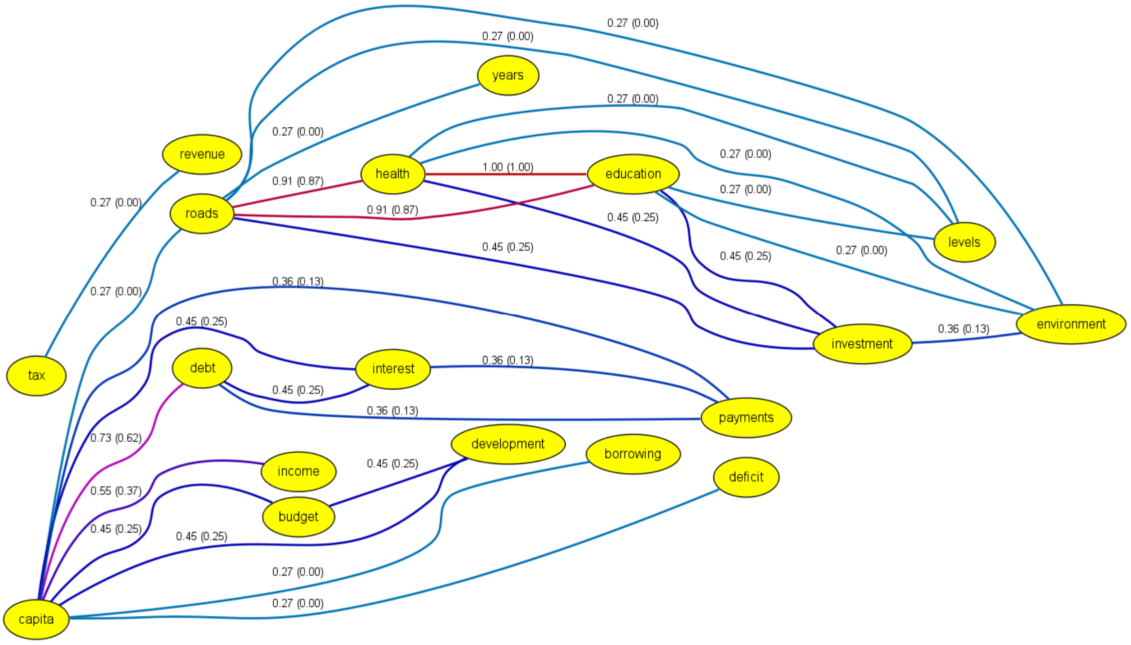


Figure 3 displays the main concepts and links used in the textual description of the optimal strategy for solving the national development planning task. The main concepts are on the vertices (nodes) of the graph. The weights from Table 2 re-appear on the links between the main concepts and measure the association strength between two concepts. Only the strongest associations are represented in the association net. In its default settings, T-MITOCAR displays up to 25 links from the list form. Therefore, there are two measures for the association strength at the links. The value outside the brackets shows the weight from the list form. The second value inside the brackets displays the weight relative to what is actually visualized. The strongest association will be 1 (“1” meaning that this is the strongest available association in the graph) and the weakest observation will be 0 (“0” meaning that this is the weakest among the strongest associations that made it into the graph).

The association net displayed in Figure 3 may be unfamiliar and thus less intuitive to system dynamicists trained in the development of stock-and-flow diagrams. An association net does not portray accumulations, feedback loops and the functional relationships described in the equations of a system dynamics simulation model. The main focus of the association net in system dynamics terms is on the structure underlying a complex dynamic problem. Process-related aspects may be represented in concepts but the behavior arising from the structure is not visible. This is, however, not the purpose of association nets which focus on the degree of association between concepts. Association nets may be able to reveal parts of the underlying reasoning heuristic when people make decisions in complex dynamic environments. In the results section of this paper we will analyze some associative (mis-)conceptions of our two dynamic decision making tasks and investigate to what degree different structural and semantic characteristics of association nets correlate with performance in dynamic decision making tasks. Most importantly for the purpose of this paper, however, association nets are the basis for the quantitative comparison of different texts.

3.1.2 Comparison of different texts and association nets

T-MITOCAR provides structural as well as semantic measures for comparing expert and non-expert texts. Such comparisons are useful in two respects:

1. They help identify misconceptions about the structure and behavior of a complex dynamic system. This can be used to increase the usability of the simulator or simulation based game for the purpose of learning, since the structure of the problem space must take the learner's epistemic belief into account (see Seel, 2003).
2. They show a progression of understanding over time (i.e., over multiple measurement time points) as compared to an expert's understanding. This is essential if a learning environment should monitor and evaluate the actual change over time (see Ifenthaler & Seel, 2005).

In order to compare texts T-MITOCAR calculates quantitative measures for structural and semantic constructs. T-MITOCAR compares association nets on a graph theoretical level. Of all the available graph theoretical measures, seven of them have shown a stable representation of different structural and semantic constructs (Ifenthaler, 2006, 2008; Pirnay-Dummer, 2006; Pirnay-Dummer, et al., 2010). The measures are based on the properties of the association net, i.e., on concepts and the links between them. The four structural and three semantic measures are defined as in Pirnay-Dummer, et al. (2010) and presented in Table 3.

Table 3: Structural and semantic measures used for the quantitative comparison of texts

	Measure	Definition
Structure	*surface* measure (see Ifenthaler, 2008)	compares the number of concepts within two graphs. It is a simple and easy way to calculate values for surface complexity
	graphical matching measure (see Ifenthaler, 2008)	compares the diameters of the spanning trees of the graphs and is an indicator for the range of conceptual knowledge.
	density of vertices measure (also often called "*gamma matching* measure") (Pirnay-Dummer, et al., 2010)	describes the quotient of concepts per concept 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 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 which state that expert knowledge is structured differently from novice knowledge).
Semantics	*concept matching* measure (Pirnay-Dummer, et al., 2010)	compares the sets of concepts within a graph to determine the use of terms. It 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.

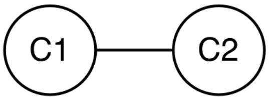
balanced semantic matching measure (see Pirnay-Dummer & Ifenthaler, 2010)	a measure which combines both propositional matching and concept matching to control for the dependency from propositional matching on concept matching: Only if concepts match, then propositions may match. BSM balances this dependency.

All structural measures quantify the comparison between the structures of written texts and all semantic measures quantify the comparison of the texts' semantics. However, the measures quantify different things and cannot substitute each other. The structural measures show convergent correlations between each other (between r=.48 and r=.79) and so do the semantic measures (between r =.68 and r =.91). The correlations between the structural and the semantic measures are lower and divergent (between r = -.24 and .36). The density of vertices (gamma) usually stands alone and only rarely correlates with the other structural measures because it accounts for a different feature of structure (correlations between r=.37 and r=.38 with the other structural measures). Pirnay-Dummer et al. (2010) provide a full validation study for the calculation of these correlation coefficients. The validation study was conducted with *N* = 1,849,926 text comparisons in 13 different subject domains ranging from common knowledge to scientific subject domains. Depending on the research questions underlying the comparison of mental models, some comparison measures may be left out. A more detailed overview of the comparison measures is provided in the following paragraphs. What every measure says qualitatively depends highly on how the text has been assessed, e.g., as to whether the compared entities have been assessed in a similar way and which task the assessment was embedded in. As with any methodology, even the best comparison measure can never level out an insufficient assessment. Thus, for models assessed with T-MITOCAR, the task needs to allow (and motivate) the subjects to reflect on the given domain within their writing. All comparison measures are conducted on graphs only, and a graph $G(V_v,E_e)$ is a set of vertices (nodes, concepts) V_v that are connected by a set of edges (links, relations) E_e. An edge connects two vertices.

3.1.2.1 Surface Matching

Within T-MITOCARs analysis, nodes are the most frequent concepts within an analyzed text, while the vertices are the most frequent syntactical associations between pair wise concepts (Figure 4).

Figure 4: A vertex links two nodes that contain concepts



The surface measure compares the number of vertices that are in a graph, to see how large a text model is.

Figure 5: Three different surface structures

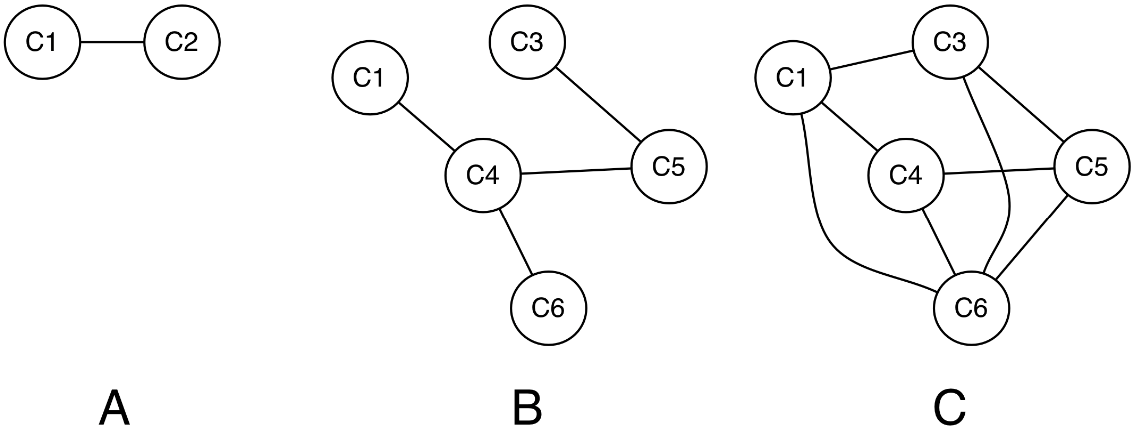


Figure 5 shows 3 different surface structures. Graph A has a surface of 1 (1 vertex), graph B has a surface of 4, and graph C has a surface of 7. The content of the nodes (concepts) does not play a part in this measure. We can now say that C is more complex than B which is more complex than A on a first superficial level. The similarity index for surface matching weights the differences between two surface measures. To calculate a similarity index between B and C, we will take the frequencies f_1 = surface (B) and f_2 = surface (C) which will result in a similarity index of

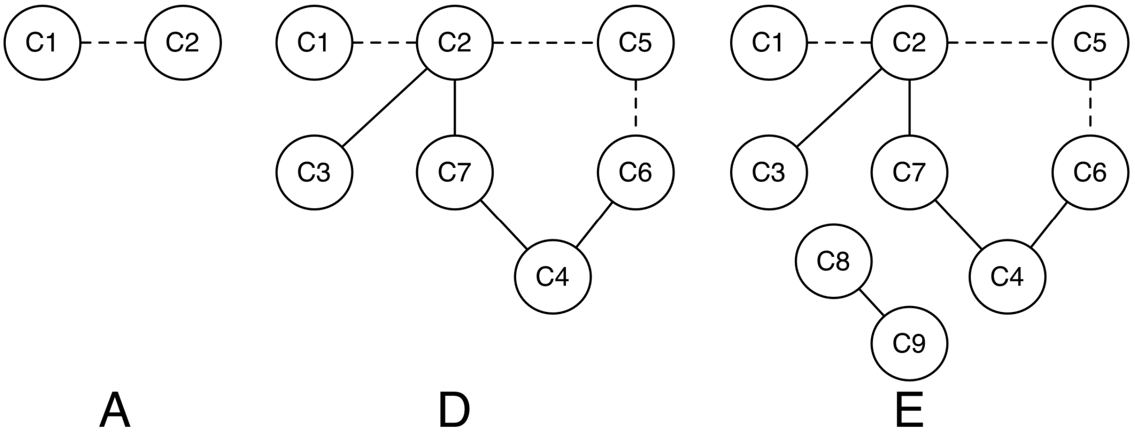
$$s_o = 1 - \frac{|f_1 - f_2|}{\max(f_1, f_2)} = 1 - \frac{7-4}{7} \approx 0{,}57$$

B and C have the same number of concepts (which may even be the same concepts) and a considerable structural similarity.

3.1.2.2 Graphical Matching

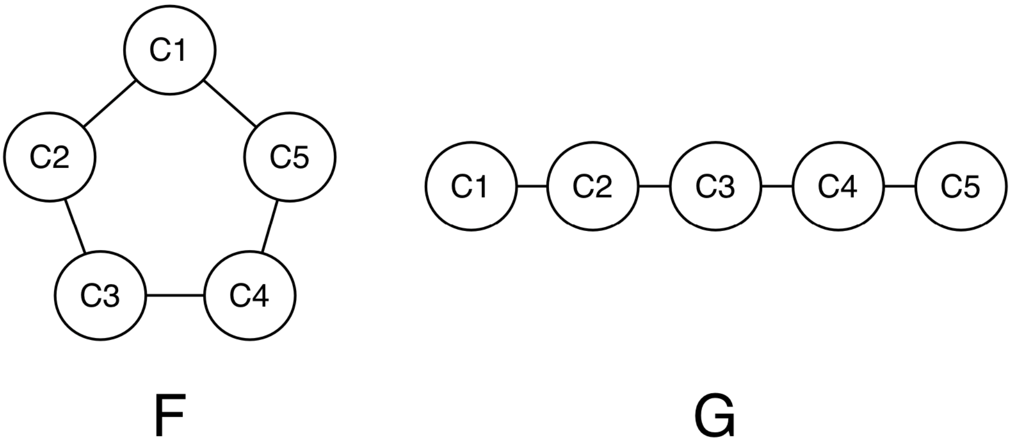
The graphical matching index takes the diameter of the spanning tree of a given graph. So, why is that and what does that mean? We start with the diameter. Between any two concepts within a graph that are at least indirectly linked to each other, there is a path.

Figure 6: Different diameters.



See Figure 6: For graph A, it takes only one step to get from C1 to C2 since there is only one vertex. The same holds for graph D when we want to go from C1 to C2. However, it takes at least 3 steps to get from C1 to C6 in graph D: C1 → C2 → C5 → C6. There is also a path that would take 4 steps, but we are interested in the *shortest* available path. Every pair of concepts that is directly or indirectly connected within a graph has such a shortest path. Pairs of concepts that are not connected do not count as can be seen from graph E: There is no connection between C9 and C6. The diameter of a graph is the longest of the available shortest paths throughout the graph. For graph A, this is obviously 1, for graph D this is 3 as it is also for graph E. If the graph has different unconnected sub graphs, the longest of the shortest paths can still be determined by the sub graph that has the longest diameter: This diameter then counts as the diameter of the whole graph. But we still need to explain the spanning tree. Diameters do have one conceptual downside when it comes to cycles. And cycles that - among other things - resemble loops play an important role within knowledge structures throughout different knowledge representations (Seel, 2003). Thus, we do not want an algorithm that is supposed to account for structure to yield half the complexity value just because a graph is or contains a cycle (a loop). On a level of the width of the graph we want the following graphs (see Figure 7) to have the same complexity value.

Figure 7: A loop and a sequence that should have the same conceptual width



The diameter alone would yield a 2 for graph F and a 4 for graph G within Figure 7 while they should at least resemble the same complexity value, maybe in some cases F would in fact be even more complex than G. But the heuristic of graphical matching cannot account for these content- and context-specific differences. A spanning tree (see Kruskal, 1957) allows for a modification of any graph that will account for that requirement. First of all, a spanning tree is a sub graph of a given graph that contains no cycles. It also connects every pair automatically with their shortest paths if possible: It finds an optimum for the distances throughout the graph.

Figure 8: Graph B is a spanning tree for graph C

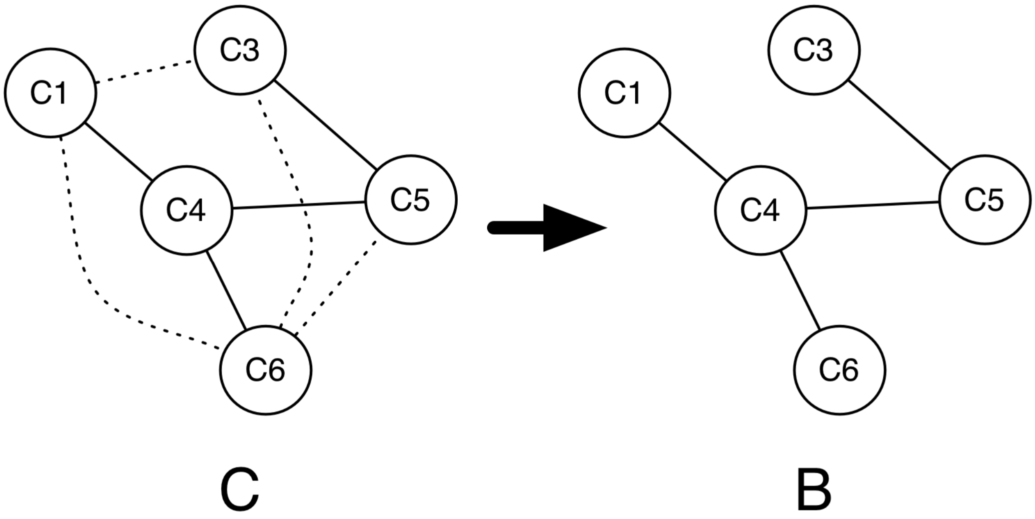


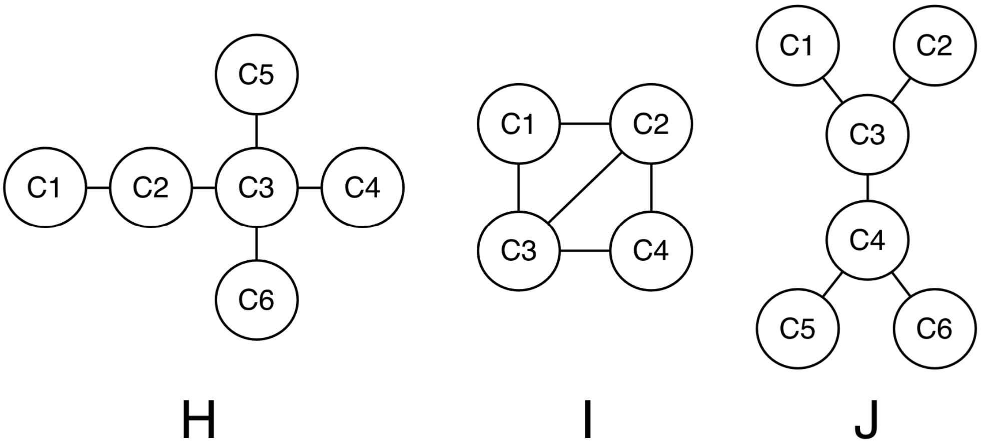
Figure 8 shows that B is a spanning tree of graph C. But why do we get rid of the cycles that may exactly have an important meaning within the knowledge model-structure? We do not apply the spanning tree on a descriptive level, but only to account for the above mentioned problem that cycles would yield half the complexity than sequences due to the nature of the diameter measure. If we now apply the diameter of the spanning tree to the graphs F and G, they are identical on the width-index of the graphical matching level; both now have a graphical width of 4 and therefore yield a similarity of 1 for the graphical matching:

$$s_G = 1 - \frac{|f_1 - f_2|}{\max(f_1, f_2)} = 1 - \frac{4-4}{4} = 1$$

3.1.2.3 Structural Matching

Structural matching is the most complex measure to derive from the graphs. It clearly has its limits in terms of computability and can thus only feasibly be applied to small and medium size graphs. T-MITOCAR generates graphs in that range, as do most concept mapping techniques when applied in experimental settings or in the usual scope of studies (see Al-Diban, 2002; Ifenthaler, 2006; Johnson, O'Connor, Pirnay-Dummer, et al., 2006; Pirnay-Dummer, 2006; Schvaneveldt, et al., 1985). At this point, only an abbreviated description of the algorithm can be provided. Please refer to Pirnay-Dummer (2010) for a detailed description and empirical testing of the structural comparison. We start with three graphs that have the same surface and graphical matching indices (Figure 9; see Pirnay-Dummer, 2010, p. 239).

Figure 9: Three graphs with the same surface and graphical matching indices.



According to the first two structural indices all three graphs (H, I, J) are identical and would yield a similarity of 1. However, they also do have structural differences that can not only be derived back to visualization differences. They have a different inner structure, which so far the two other indices do not account for. Differences are easy to recognize by human viewers when the graph is visualized, but not as easily quantifiable. Structural matching uses an analytical approach that is automatable. As a first step, a graph is split into several different possible sub graph-pieces. The resulting pieces are not exclusive, so that the results become quite complex even for the simple graph I (Figure 10).

Figure 10: Basic sub graphs and their frequencies within graph I from Figure 9

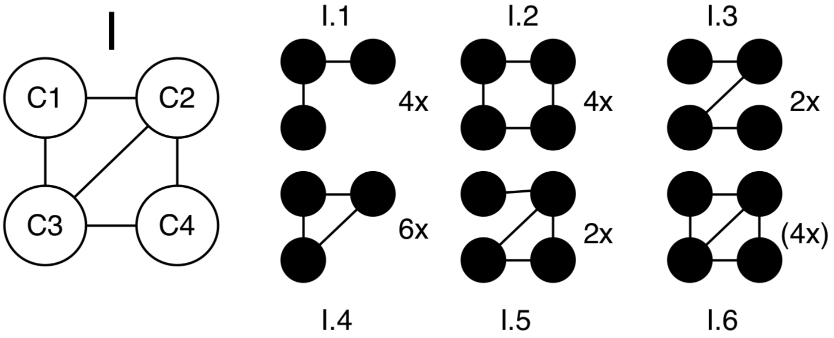


Figure 10 shows the sub graphs and the frequency of their occurrence as constructed by the structural comparison measure. Essentially, from every node, every possible path through the graph is constructed as a sub graph. These sub graphs are called “traces”. Cycles (loops) are treated in a special way. On the one hand, as a stop criterion, path reconstruction has to be halted once a node is passed twice. On the other hand, this gives an opportunity to specially code a cycle. Paths are reconstructed from each member of the cycle in the same way. Therefore the node occurs more often than in non-cyclic structures, e.g., in I.4 within Figure 10: the basic form occurs only twice but due to the cycle it will be constructed for each member thus leaving $f_{I.4} = 2 \cdot 3 = 6$ traces of the same basic form. The traces are usually not visualized during analysis; they are instead represented on strings (Table 4). Depending on the graph’s complexity, there can be millions of possible traces.

Table 4: String representation for the traces in graph I (Figure 10)

Sub graph	Trace-String	Frequency
I.1	1,2,1	4
I.2	2,2,2,2,-2	4
I.3	1,2,2,1	2
I.4	2,2,2,-2	6
I.5	1,3,2,2,-3	2
I.6.	3,2,3,2,-3	2·2=4
	2,3,2,3,-2	
Σ		22

Table 4 shows how the traces are represented on strings. For each node that the trace passes, the string contains the number of vertices (or in graph theory terms: the degree) that this node has as:

$$\Gamma(\nu) = 2$$

A whole string is then an ordered chain of degrees:

$$\Gamma_i(\nu) = \{1,3,2,2,-3\}$$

A complete cycle will be encoded with a “-“ at the last node.

In order to retrieve all underlying basic shapes automatically, the graph needs to be processed in two steps. First, all paths that are directly available from the graph’s vertices V_v need to be collected in a set for each possible path length u:

$$\Gamma^{V}_{\upsilon,i}(V_v) = \bigcup_{i=1}^{n}\bigcup_{u=2}^{\upsilon}\Gamma_{u,i}(\nu) = \{\{1,3,2,2,-3\},\{1,2,1\},...\}$$

There is also a stop criterion υ as a maximum length (range) for the traces that are constructed. With that stop criterion the algorithm focuses only on the basic traces and balances computability at the same time. Traces beyond that range will not be constructed. Once all directly accessible traces are constructed, implicit traces are added. One of these traces was already “illegally” illustrated in Figure 10, I.2: When the algorithm traces through the graph, this shape will at first not occur: There is no node configuration that directly leads to {2,2,2,2,-2}, the only traces that the algorithm will find are {2,3,2,3,-2} twice and {3,2,3,2,-3} twice. Therefore also the underlying sub traces are built, i.e., {2,2,2,2,-2}. This second procedure is called *downtrace*. The final set contains all directly available sets plus all the possible downtraces thereof. To indicate that a downtrace has been constructed, we write the downtrace as function Ξ on all the directly available traces:

$$\Xi\left(\Gamma^{V}_{\upsilon,i}(V_v)\right)$$

Although the number of downtraces has shown to be highly selective between experts and non-experts (see Pirnay-Dummer, 2010), it is usually not taken into account when structural similarity is calculated. Once all downtraces are available for two graphs, then the Tversky Similarity Measure (Tversky, 1977) is applied to compare the two sets of downtraces:

$$s_{\Xi} = \frac{f(\Xi_A \cap \Xi_B)}{f(\Xi_A \cap \Xi_B) + \alpha \cdot f(\Xi_A - \Xi_B) + \beta \cdot f(\Xi_B - \Xi_A)}$$

Intersection and difference sets are both incorporated, and α and β can be used to balance the difference sets if the models of the graphs had been assessed in a completely different way and would therefore yield unfair comparisons, e.g., when an expert had more time than the subjects to create their expert model. However, we strongly suggest to control for this on the methodological side whenever possible and keep α and β equally weighted as α =β=.5. With all that at hand, the structural matching indices between the graphs H, I, and J from Figure 9 can be calculated as shown in Table 5. The structural matching index reveals the differences between the graphs.

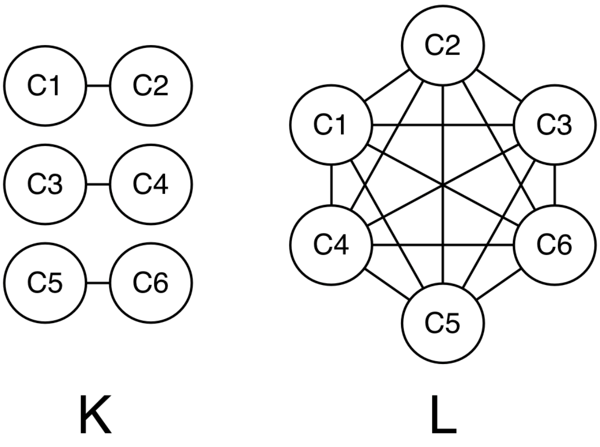
Table 5: Structural matching measures between the graphs H, I, and J from Figure 9

	J	I
H	0.71	0.32
J		0.47

3.1.2.4 Gamma Matching

The gamma measure stands for an internal connectedness. Its raw value is simply calculated as a relative measure of links per concept, and it thus projects a density of vertices. The procedure and reason is easily understandable when illustrated with the help of two structurally weak models (Figure 11).

Figure 11: Two structurally weak models on a graph.



Graph K from Figure 11 shows only single pair wise links, nothing is really interconnected, whereas graph L just connects everything with everything. The latter usually yields higher plausibility, because untrained viewers can always find their specific ideas represented. Both represent the possible extremes to each side. K has a raw value of 2 nodes per link (6 terms and 3 links), and L has one of 0.4 (6 terms, 15 links). The top raw value is always 2 but the bottom value depends on the number of concepts. Thus, the raw value Rγ is:

$$\frac{n}{\frac{n!}{2\cdot(n-2)!}} \le R_\gamma \le 2$$

To produce a readable output, the raw value is linearly scaled as γ between 0 (resembling the bottom value) and 1 (resembling the top value, i.e., 2). Several studies showed that expert models seem to have a gamma value of around $\gamma = 0.35$ (Pirnay-Dummer, 2006). However, also in these studies, the gamma value was not very selective between experts and non-experts. A final conclusion about the quantifying discriminatory power of the gamma matching is still a pending research question. After retrieving the individual values for gamma, its matching index between two graphs can be calculated as:

$$s = 1 - \frac{|f_1 - f_2|}{\max(f_1, f_2)} = 1 - \frac{1-0}{1} = 0$$

Of course, for the example graphs in Figure 11 this yields no similarity and a gamma matching of 0.

3.1.2.5 Concept Matching

Concepts within a graph are semantically represented by concrete terms. A term does not necessarily constitute the presence of a concept per se. A term can be an instance of a concept. However, once a term is embedded frequently in a similar way into its neighboring web, it is likely that the term fills a concept. Due to its embeddings we heuristically assume, that if a term that converges into a net that has sufficient stability (frequency) then it would also resemble a concept. Concept matching is the first semantic measure. Semantic measures look at the content as opposed to the structural measures that look at how the graph is structurally composed. As the name suggests, the first measure aims at the concepts that are used within a model.

Figure 12: Two different concept maps from music.

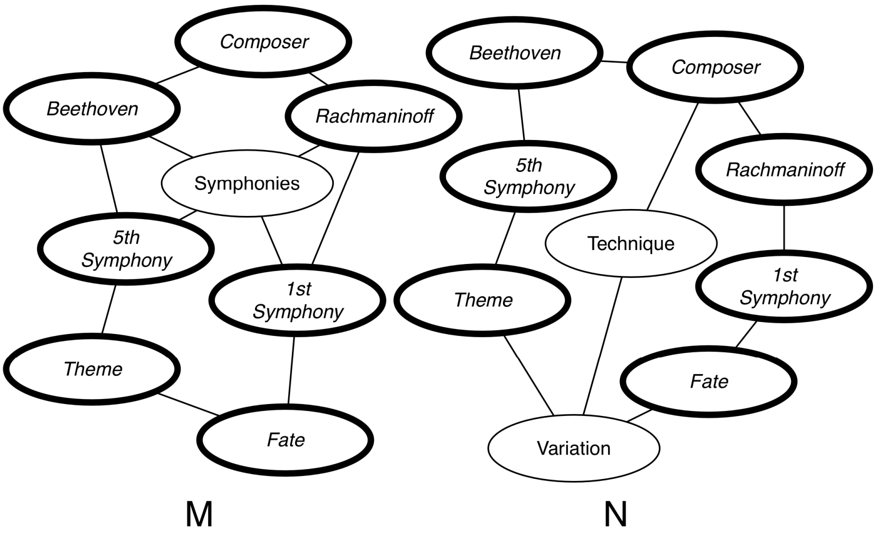


Figure 12 shows two graphical models that represent similar aspects from music. Both contain two specific composers and focus on the connection of two of their symphonies, namely Rachmaninoff’s fate theme that is composed as a variation on a theme from

Beethoven's 5th symphony. M is more hierarchical and points out the symphonies as a central concept. It has the fate theme but leaves it open as to which musical piece this belongs to. N focuses on the variation technique as a common composing method, thus also showing that the fate theme is a variation on another symphonies theme. We leave the interpretation as to which model should be considered more expert-like to the experts themselves, consider them both as different learners' models, and look into the language use, particularly the use of concepts between both text models. The matching concepts are already marked in both graphs. We have a set that matches (7 concepts) and two difference sets (1 concept in model M and 2 concepts in model N) for each graph. Again, Tverksy-Similarity is used to calculate the concept matching measure:

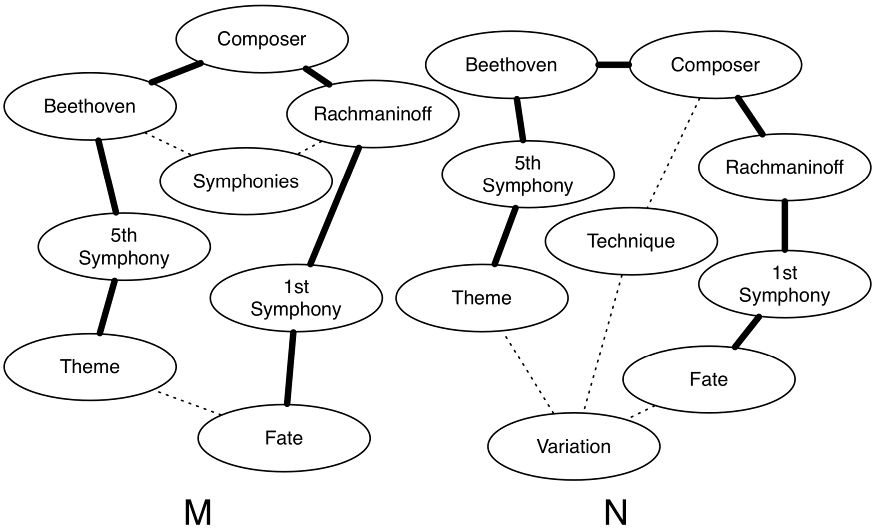
$$s_v = \frac{f(v_M \cap v_N)}{f(v_M \cap v_N) + \alpha \cdot f(v_M - v_N) + \beta \cdot f(v_N - v_M)} = \frac{7}{7 + .5 \cdot 1 + .5 \cdot 2} \approx .82$$

As for structural matching α and β can be used to balance the difference sets, but again we recommend to control for this already by the assessment methods and leave the two at .5 each. In our case, N and M match as regards their concepts with a similarity of s=.82.

3.1.2.6 Propositional Matching

A real proposition needs to specify its predicate that connects two concepts. To that end the propositional matching does not use full propositions. It heuristically assumes that if two concepts are stably connected that there is an inherent reason for this association. Thus, propositional matching can be calculated with or without knowing the annotations at the links, e.g., equations, causal directions, or hierarchical functions. Propositional matching does something similar to concept matching. It counts the match of the edges (links, relations) between two graphs. We use the same graph as for concept matching, but with a different focus to illustrate the measure (Figure 13).

Figure 13: Different propositions match between the two graphs N and M.



7 propositions match between graph N and M (Figure 13), M has 3 propositions that do not match (difference set) and N has 4. Again, Tversky Similarity is calculated to determine the propositional matching measure.

$$s_e = \frac{f(e_M \cap e_N)}{f(e_M \cap e_N) + \alpha \cdot f(e_M - e_N) + \beta \cdot f(e_N - e_M)} = \frac{7}{7 + .5 \cdot 3 + .5 \cdot 4} \approx .67$$

The concepts may be very well aligned and the models may have semantic similarities on the propositions. However, we also see that they are about something else in some details of the models.

3.1.2.7 Balanced Semantic Matching

This measure uses a combination of concept matching and propositional matching to balance for a specific dependency between both semantic measures. It is however not an aggregation of them and should not be misinterpreted in that way. Propositional matching is dependent on concept matching: Only if concepts match then also propositions can match. The more concepts match, the more propositions *may* match. But propositions obviously do not automatically match only because the concepts do. Sometimes this dependency is hard to interpret from the first two semantic measures alone, especially when individual comparisons are aggregated otherwise (e.g., within group means). Balanced semantic matching accounts for this dependency by dividing propositional matching by concept matching, except if no concepts match then also balanced semantic matching is 0.

$$s_{bsm} = \begin{cases} \frac{s_e}{s_v}, s_v > 0 \\ 0, else \end{cases} = \frac{.67}{.82} \approx .81$$

Thus, the balanced semantic matching value between the graphs N and M is s = .81.

3.2 Manual analysis

The purpose of the manual analysis was to validate the automated analysis with T-MITOCAR for the application to complex dynamic problems. We only compared the two analysis methods for the national development planning task as a manual analysis of the verbal descriptions of the 129 subjects in the reindeer rangeland management task was virtually impossible.

Subjects' understanding was assessed using the questions about their description of the problem situation and the proposed solution strategy, which had been asked both before and after using the simulation. The responses of the subjects were printed on one side of an index card and their subject number was on the reverse side to enable blind scoring. A scoring protocol was devised that roughly assessed subjects' understanding of detail complexity and dynamic complexity (Senge, 1990). Detail complexity represents the amount of content and can be measured for example, by the number of variables or concepts and the number of links between them. Dynamic complexity refers to the presence of feedback thinking and appreciation of other important system dynamic concepts such as delays and nonlinearities. The scoring protocol awarded a number of points to these elements with the maximum number of points determined by the expert text.

In the expert text, we identified 16 relationships between important variables. An example of such a relationship is that per capita income depends on capital and total factor productivity. Subjects received one point for each of those relationships that they identified, the maximum being 16.

To measure subjects' understanding of stock and flow variables and their interactions, points were assigned if subjects were able to infer the characteristics of successful investment strategies. In total we coded the verbal descriptions for a maximum of 6 such characteristics. Subjects received one point if their description included the concept of balancing education, health and roads (recognition of non-linearities). 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 should begin to 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.

4 Results

The comparison of subjects' verbal descriptions of the problem situation and their intended strategy to solve a dynamic decision making task and the expert descriptions helps revealing misperceptions of structure-behavior relationships in the decision making task (section 4.1). If subject texts are available for several measurement time points during the decision making task (e.g., before, while and after interacting with a simulator or a simulation based game) a progression of improved understanding over time (using an expert's understanding as a standard) can be provided. Similarly, the differences in understanding between different experimental conditions (e.g., experimental and control group) can be analyzed (section 4.2). Verbal descriptions of a problem situation and the intended strategy to solve the problem can also be compared to the performance measures recorded during the decision making trials. With such comparisons, the relationships between elements of understanding and performance can be analyzed (section 4.3). For the national development planning task we also compare the results from the automated and the manual analysis methods to gain insight into the validity of the automated analysis in the context of dynamic decision making tasks.

The vast majority of the subjects in the national development planning tasks were students for whom English is a second language. Their problem and strategy descriptions were, to different degrees, lexically or grammatically incorrect or confusing for the automated scoring program. After an initial attempt at completely automated analysis, we filtered and fixed the protocols by hand. This could be done fairly objectively and thus should not distort the results. For the national development planning task we analyzed the data both for a long expert text and for a shorter expert text that contained less technical jargon and only a summarized description of the problem situation and solution strategy (see appendix 3). We only report on results with the long expert text as the results for the long and short expert texts were the same.

4.1 Comparing non-expert to expert texts

4.1.1 Automated analysis

For a quantitative comparison of subject and expert texts T-MITOCAR automatically generates association nets and calculates the structural and semantic indices described in the methods section. From 377 descriptions of the problem situation and the proposed strategy to solve the reindeer rangeland management task 365 could be automatically represented as an association net. The 12 texts that could not be represented as association nets were too short for T-MITOCAR to analyze in a meaningful way. All of the 78 descriptions of the problem situation and the proposed strategy to solve the national development planning task could be automatically represented as an association net.

Table 6 contains the comparison measures that T-MITOCAR calculated for the two dynamic decision making tasks. The measures indicate the similarity between the average subject and the expert for texts provided by the subjects after reading the instructions but before interacting with any computer simulation tools. A value of 1 for any of the indices in the table would indicate that the subject text is equal to the expert text for a specific structural or semantic characteristic.

Table 6: Structural and semantic indices between the expert's and the subjects' strategy description in the two dynamic decision making tasks

		Reindeer rangeland management task	National development planning task
	Index	Similarity, mean (SD)	Similarity, mean (SD)
Structure	Surface Matching	0.51 (0.26)	0.23 (0.07)
	Graphical Matching	0.66 (0.22)	0.40 (0.13)
	Structural Matching	0.58 (0.31)	0.20 (0.16)
	Gamma Matching	0.62 (0.24)	0.50 (0.36)
Semantics	Concept Matching	0.38 (0.13)	0.31 (0.09)
	Propositional Matching	0.14 (0.10)	0.11 (0.07)
	Balanced Semantic Matching	0.38 (0.96)	0.76 (0.51)

Table 6 indicates that for both decision making tasks, the structure of the subjects' texts is closer to that of the experts than is the semantics of the texts. As the overall structural similarities are high, this may be interpreted as in indicator that the method of reasoning may point in the right direction. As the semantic similarities are low, it might however be that the subjects focus on irrelevant features of the problem situation and the solution strategy. Low semantic similarities indicate that subjects use different concepts and different propositions than the expert text. Individual subject association nets for the reindeer rangeland management task, for example, mention the relevant stock (lichen) but fail to mention the flows that change the stock, lichen growth and grazing (consumption) by the reindeer. The failure to pay attention to the flows makes it difficult for subjects to make effective decisions.

Cronbach's standardized alpha reliability was α=.61 (.73) between the three semantic measures and α=.89 (.90) between the structural measures for the reindeer rangeland management task (national development planning task). The reliability measures indicate the degree to which T-MITOCAR is able to generate stable results. The measures are slightly lower than in general validation studies (e.g., Pirnay-Dummer & Ifenthaler, 2010) which implies a moderate effect that is specific to dynamic decision making tasks.

4.1.2 Manual analysis

Table 7 lists the relationships and characteristics of successful strategies described in the expert text for the national development planning task. The last column of the table indicates the number of times these relationships or characteristics were mentioned in the subject texts.

Table 7: Frequency of relationships and characteristics of successful strategies described in the subject texts for the national development planning task

relationships	goal: max. pc income-interest payments	13
	pc income = f(capital, TFP)	10
	capital increases with investment	2
	investment increases with pc income	1
	investment increases with education	10
	investment increases with health	10
	investment increases with roads	10
	PM can regulate resource expenditure	16
	available budget =tax revenue-interest payments	5
	tax revenue = pc income * tax rate	16
	deficit when desired > available budget	2
	surplus when desired < available budget	1
	deficit leads to borrowing	9
	borrowing leads to debt	14
	debt leads to interest payments	13
	surplus leads to paying down	3
strategy	balance resources	3
	education early	13
	roads early	12
	health later	4
	borrow early	13
	pay down later	3

Table 7 illustrates that a high number of subjects were able to identify the key stocks in the system (capital, education, health, roads, and debt). While they are able to see the capital stock (“pc income =f(capital, total factor productivity)” relationship) they fail to mention the inflow (capital increases with investment). Similarly, many subjects describe the debt stock and that it increases with borrowing. Only very few, however, are also able to describe the outflow that can decrease the stock (that is, the “surplus leads to paying down” relationship).

Only one subject is able to close the private sector development loop between capital, pc income, investment and capital. Also, very few subjects describe the budget mechanisms correctly (that is, the “deficit when desired > available budget” and the “surplus when desired < available budget” relationships).

The missing focus on the flows is confirmed in the descriptions of the characteristics of successful strategies. Many subjects realize that they need to finance the important early investments in education (because of the long delay) and roads (because of the rather immediate impact on growth) through borrowing. However, only very few subjects mention the importance of paying down debt later to avoid exponentially growing interest payments on debt.

The majority of subjects fail to recognize the importance of the non-linearities in the system, with only 3 subjects mentioning that the three resources (education, health and roads) need to be balanced for maximum growth and that as a result, investment in health must be increased a bit after investment increases in education and roads.

4.2 Comparing texts over time and across groups

4.2.1 Automated analysis

Table 8 lists the means for the structural and semantic indices at three different measurement time points (before, while and after interacting with the simulation tools) and for two experimental treatments (control group, experimental group using a beneficial learning strategy) for the reindeer rangeland management task. None of the differences between the measurement time points and treatments are statistically significant, indicating that understanding did not increase over time or due to a specific experimental condition.

Table 8: Structural and semantic indices (means) between the expert's and the subjects' problem and strategy description for the treatments and the measurement time points in the reindeer rangeland management task

		MTP.1		MTP.2		MTP.3	
	Index	Ctr	Exp	Ctr	Exp	Crt	Exp
Structure	Surface Matching	0.60	0.60	0.60	0.60	0.62	0.59
	Graphical Matching	0.72	0.71	0.68	0.71	0.69	0.72
	Structural Matching	0.63	0.63	0.64	0.60	0.64	0.60
	Gamma Matching	0.61	0.62	0.59	0.61	0.60	0.63
Semantics	Concept Matching	0.38	0.39	0.40	0.40	0.41	0.41
	Propositional Matching	0.14	0.16	0.15	0.16	0.15	0.16
	Balanced Semantic Matching	0.25	0.40	0.24	0.34	0.23	0.37

Ctr = Control Group; Exp = Experimental Group; MTP = Measurement Time Point

In the case of the national development planning task there were also no significant differences in the indices between the measurement time points (MTP; before and after interacting with simulation tools). However, the results displayed in Table 9 show significant differences between the control group and the experimental group for the structural indices. The experimental group has higher similarity indices than the control group. However, the two groups already differ at the first measurement time point, i.e., right after studying the instructions to the task but before interacting with the simulation. The better indices for the experimental group can therefore not entirely be explained by the beneficial instructional strategy.

Table 9: Structural and semantic indices (means) between the expert's and the subjects' strategy description for the treatments and the measurement time points in the national development planning task

		MTP1			MTP2		
	Index	Ctr	Exp	sig.	Ctr	Exp	sig.
Structure	Surface Matching	0.38	0.40	*	0.34	0.40	*
	Graphical Matching	0.56	0.61	**	0.64	0.58	**
	Structural Matching	0.44	0.53	*	0.43	0.54	**
	Gamma Matching	0.63	0.57		0.64	0.53	
Semantics	Concept Matching	0.28	0.33		0.29	0.34	
	Propositional Matching	0.09	0.06		0.09	0.06	
	Balanced Semantic Matching	0.18	0.37		0.17	0.41	

Ctr = Control Group; Exp = Experimental Group; MTP = Measurement Time Point; sig. = significance

* significant at .1; ** significant at .05; *** significant at .001

According to Table 9 the experimental subjects used a significantly higher number of concepts in their textual descriptions than the control subjects (surface matching), the range of the concepts was significantly larger (graphical matching) and the entire structure of their association nets is significantly closer to the structure of the expert text (structural matching). It is interesting to note that the similarity in the case of the surface and the graphical matching decline slightly after interacting with the simulation tools (indices at measurement time point 2 (MTP2) < indices MTP1). The changes are, however, not statistically significant. The values for the structural matching increase from measurement time point one to measurement time point to two indicating a shift towards a more expert structure of the text.

4.2.2 Manual analysis

Table 10 lists the mean number of relationships and characteristics of successful planning strategies described by the subjects of the national development planning task. Similar to the automated analysis the table differentiates between measurement time points (before and after interacting with the simulation tool) and two experimental conditions.

Table 10: Number of relationships and characteristics of successful planning strategies for the treatments and the measurement time points in the national development planning task

	MTP.1			MTP.2		
	Ctr	Exp	sig.	Ctr	Exp	sig.
Relationships	3.9	4.1		3.9	4.3	
Strategy	0.9	1.8	**	0.9	2.1	***

Ctr = Control Group; Exp = Experimental Group; MTP = Measurement Time Point; sig. = significance

* significant at .1; ** significant at .05; *** significant at .001

The two experimental conditions as well as the two measurement time points are not different in terms of the mean number of relationships described in the subject texts. The experimental conditions, however, have significant differences in the number of characteristics of effective or successful strategies. As is the case in the automated analysis, the experimental group is already better than the control group at measurement time point one, i.e., right after studying the instructions for the task. The differences between the two groups become bigger at measurement time point two where the experimental group describes even more characteristics of successful strategies.

4.3 Understanding of and performance in dynamic decision making tasks

The performance results of the reindeer management task did not show significant differences for the measurement time points or experimental conditions (experimental versus control group). We were also not able to detect any significant relationships between the structure and semantic measures and performance in the three decision making trials.

Table 11 lists the correlation coefficients between performance in the national development planning task (value of per capita income minus per capita interest payments on debt) and the seven understanding measures generated by T-MITOCAR. Figure 14 illustrates graphically (to emphasize changes) the development of the correlation between performance and the understanding measures (similarity measures) over time.

Table 11: Correlation coefficients between performance and understanding measures in the national development planning task

		Structure				Semantics		
	Year	Surface Matching	Graphical Matching	Structural Matching	Gamma Matching	Concept Matching	Propositional Matching	Balanced Semantic Matching
Trial 1	2015	-0.11	-0.09	-0.04	0.18	-0.01	-0.16	-0.07
	2020	-0.13	-0.13	-0.08	0.13	0.01	-0.14	-0.05
	2025	-0.10	-0.08	-0.04	0.08	0.03	-0.11	-0.03
	2030	0.07	0.11	0.11	0.17	0.11	0.04	0.05
	2035	0.20	0.24	0.22	0.21	0.15	0.13	0.08
	2040	0.26	0.31	0.27	0.22	0.16	0.17	0.09
	2045	0.28	0.33	0.29	0.23	0.17	0.19	0.10
	2050	0.29	0.35	0.30	0.23	0.17	0.20	0.10
	2055	0.29	0.35	0.30	0.23	0.17	0.20	0.10
	2060	0.29	0.35	0.30	0.23	0.17	0.20	0.10

Figure 14: Performance similarity fit for the national development planning task

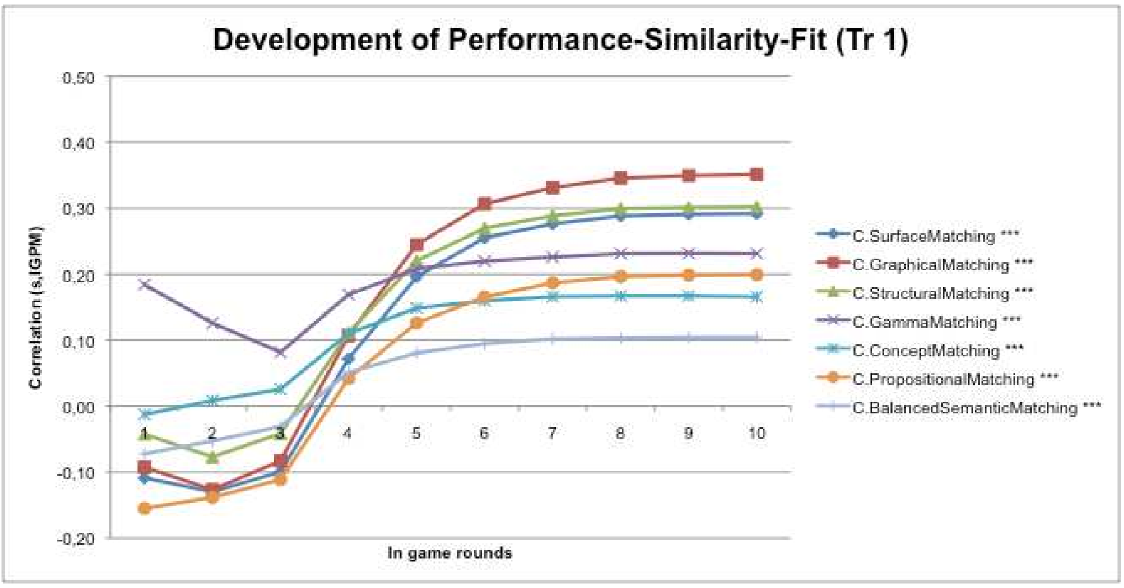


Figure 14 shows the convergence of the correlations between the performance measures (in-game-performance-measures IGPM) and the similarity measures (s) over time during the first trial. *** indicate correlations that are statistically significant, p<.001). The “in game rounds” label of the x axis denotes the time in the simulation model with 1 representing the year 2015 and 10 the year 2060. The correlation coefficients increase over time. The more similar a subject text becomes with respect to the expert text (i.e., the higher its similarity measure) the better performance becomes at the end of the game. From this we can conclude that the text descriptions (represented by the understanding or similarity measures) can explain the performance measures better over time. Both structural and semantic measures converge to the performance measures and can more stably explain the resulting performance. In general, the structural measures show higher correlation coefficients than the semantic measures. Understanding of different aspects of the structure of the complex dynamic problem therefore determines performance more than understanding of entire propositions (concept-link-concept) does.

5 Discussion

In our ongoing research on learning in system-dynamics based learning environments, evaluating and improving such systems depends foremost on our ability to measure the learners’ outcomes. Those outcomes are of two main categories: learners’ performance within the environment (to what extent they make decisions which result in beneficial simulation results) and learners’ understanding (to what extent they correctly comprehend the nature of the underlying simulation model and the principles for effectively managing the simulation model). Measuring performance within the learning environment is relatively straightforward. Measuring understanding, which is inside the learners’ heads, is much more difficult. In previous work we have used the traditional method of asking learners to write statements describing their understanding and thinking processes. Those verbal protocols are then analyzed by human raters who search for statements in the protocols which reflect either good or poor understanding and good or poor thinking, such as a strategy for managing the simulation. Such analysis must be done very systematically and objectively, for example, by raters being blind to the experimental conditions represented by particular protocols. Such “by hand” analysis is difficult and time-consuming, and their own validity issues are hidden by their methodological nature: Even a sufficiently available objectivity does not necessarily lead to validity. This study subjected data from research participants to an automated analysis (T-MITOCAR) which compares learners’ protocols to expert protocols, and did so across time (to assess learners’ improvement) and between experimental conditions (to assess the relative effectiveness of different instructional strategies within a system dynamics based learning environment).

Our main research goal was to evaluate the validity between the manual and the automated quantitative method for assessing understanding in dynamic decision making tasks. Our past work, using manual analysis of verbal protocols, has demonstrated significant differences in understanding between learners using beneficial learning strategies (preliminary behavior exploration) and learners not using such strategies (Kopainsky & Sawicka, accepted; Kopainsky, et al., 2009). The automated analysis would demonstrate validity for measuring learner understanding if its numerical indices showed similarly that learners working with good instructional strategies did better than learners with poorer (or more traditional) instructional strategies. Our analysis of the automated indices tends towards supporting its validity in the case of the national development planning task. The manual as well as the automated analysis revealed a missing focus of the subjects’ texts on the flows that can change the stocks. Both methods for analyzing differences between experimental conditions found that learners receiving the theoretically better instructional strategy improved more (from their initial description to their final description) than did the learners receiving the theoretically poorer instructional strategy. The results from the automated analysis are thus in keeping with our theoretical prediction and in line with the manual analysis, which provides a form of construct validity.

While the validity in general seems to be given a number of issues have to be considered. First, the reindeer rangeland management task revealed no significant results for subject-expert comparisons or for learner progression over time or comparisons between experimental conditions. The system dynamics model underlying the reindeer rangeland management task is a very small model with only one stock and two flows. Verbal descriptions of both the model structure and the optimal strategy to solve the task can only be very short. Thus, minor differences in the verbal descriptions and the use of terms may result in major deviations from the expert text. An automated analysis of verbal protocols such as the one by T-MITOCAR therefore seems to be practical only for tasks with larger underlying simulation models.

Second, the number of participants in the national development planning task was rather low and thus the statistical power of our correlation analyses could be improved. In addition to more subjects, modifications in the decision making task itself might help identifying the understanding related drivers of performance in more detail. As discussed in past studies (e.g., Pirnay-Dummer, 2006), a good writing task is very important for the analysis to work. Such a task would not only induce the learners to write essentially more text, but to be more precise about what they write. There are several known ways to construct better writing tasks. Task embeddedness is one approach, in which the writing becomes an integral part of the task itself to make it less obvious that it is used for assessment. Usually, to write for a peer or to write, for example, to a "local politician in order to help him or her with a certain problem" has shown to be successful.

Although we believe that system dynamics-based learning environments hold great potential for improving important types of human performance, attaining such improvement depends upon two things: designing those environments and conducting research to measure their learning effectiveness. Measuring learning effectiveness has always required great time and effort. As a part of our program to design and improve system dynamics-based learning environments, better methods to measure learning effectiveness are necessary. In this study we have begun the investigation of automated methods for assessment of learning outcomes. Success with them will enhance our capacity to evaluate and improve the learning environments we design. The results of this study provide some evidence that automated analysis of learning outcomes can be as accurate and valid as more traditional and labor-intensive methods. Our next step is to refine our collection of learning data (e.g., using embedded story questions) for input to the automated technique, and to validate that for different conditions in which differences in learning can be expected to occur.

6 Acknowledgements

We would like to thank Krystyna Stave for giving us access to her student pool for the reindeer rangeland management task. Pål I. Davidsen and Matteo Pedercini helped collect data for the national development planning task. One of the authors was supported by the Swiss National Science Foundation under a Fellowship for Advanced Researchers. The views and conclusions expressed in this paper are those of the authors alone and do not necessarily reflect the views of the Swiss National Science Foundation.

7 References

Al-Diban, S. (2002). *Diagnose mentaler Modelle*. Hamburg: Kovac.

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

Capelo, C., & Dias, J. F. (2009). A system dynamics-based simulation experiment for testing mental model and performance effects of using the balanced scorecard. *System Dynamics Review, 25*(1), 1-34.

Cavaleri, S., & Sterman, J. D. (1997). Towards evaluation of systems thinking interventions: a case study. *System Dynamics Review, 13*(2), 171-186.

Cavaleri, S. A., & Thompson, J. P. (1996). *Assessing the efficacy of microworlds for promoting systems thinking.* Paper presented at the International System Dynamics Conference, Cambridge, MA.

Diehl, E., & Sterman, J. D. (1995). Effects of feedback complexity on dynamic decision making. *Organizational Behavior and Human Decision Processes, 62*(2), 198-215.

Doyle, J. K. (1997). The cognitive psychology of systems thinking. *System Dynamics Review, 13*(3), 253-265.

Doyle, J. K., Radzicki, M. J., & Trees, W. S. (2008). Measuring change in mental models of complex dynamic systems. In H. Qudrat-Ullah, J. M. Spector & P. I.

Davidsen (Eds.), *Complex Decision Making* (pp. 269-294). Berlin/Heidelberg: Springer.

Hanke, U. (2006). *Externale Modellbildung als Hilfe bei der Informationsverarbeitung und beim Lernen.* Freiburg: Universität Freiburg. Institut für Erziehungswissenschaft.

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.

Huz, S., Andersen, D. F., Richardson, G. P., & Boothroyd, R. (1997). A framework for evaluating systems thinking interventions: An experimental approach to mental health system change. *System Dynamics Review, 13*(2), 149-169.

Ifenthaler, D. (2006). *Diagnose lernabhängiger Veränderung mentaler Modelle Entwicklung der SMD-Technologie als methodologisches Verfahren zur relationalen, strukturellen und semantischen Analyse individueller Modellkonstruktionen*. Freiburg: FreiDok.

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

Ifenthaler, D., Masduki, I., & Seel, N. M. (2009). The mystery of cognitive structure and how we can detect it. Tracking the development of cognitive structures over time, *Instructional Science*.

Ifenthaler, D., & Seel, N. M. (2005). The measurement of change: Learning-dependent progression of mental models. *Technology, Instruction, Cognition and Learning, 2(4)*, 317-336.

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

Jensen, E., & Brehmer, B. (2003). Understanding and control of a simple dynamic system. *System Dynamics Review, 19*(2), 119-137.

Jensen, E., & Sawicka, A. (2006, July 23-27). *What is the use of basic dynamic tasks?* Paper presented at the 24th International Conference of the System Dynamics Society, Nijmegen, The Netherlands.

Johnson, T. E., O'Connor, D. L., Pirnay-Dummer, P., Ifenthaler, D., Spector, J. M., & Seel, N. M. (2006). Comparative study of mental model research methods: Relationships, among, ACSMM, SMD, MITOCAR & DEEP methodologies. In A. J. Cañas & J. D. Novak (Eds.), *Proceedings of the Second International Conference on Concept Mapping*. San Jose, Costa Rica: Universidad de Costa Rica.

Johnson, T. E., O'Connor, D. L., Spector, J. M., Ifenthaler, D., & Pirnay-Dummer, P. (2006). Comparative study of mental model research methods: Relationships among ACSMM, SMD, MITOCAR & DEEP methodologies. In A. J. Canas & J. D. Novak (Eds.), *Concept Maps: Theory, Methodology, Technology*. San Jose, Costa Rica: Proceedings of the Second International Conference on Concept Mapping.

Johnson-Laird, P. N. (1983). *Mental Models. Toward a cognitive science of language, inference and language*. Cambridge: Cambridge Univ. Press.

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development, 48*(4), 63-85.

Just, M. A., & Carpenter, P. A. (1976). The relation between comprehending and remembering some complex sentences. *Memory and cognition, 4*(3), 318-322.

Klein, G. (1997). The recognition primed decision model:looking back, looking forward. In C. E. Zsambok & G. A. Klein (Eds.), *Naturalistic decision making*: Lawrence Erlbaum Associates.

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

Kruskal, J. B. (1957). On the shortest spanning subtree of a graph and the travelling salesman problem. *Proc. American Math. Society 7,* 48-50.

Maani, K. E., & Maharaj, V. (2001). *Systemic thinking and problem solving: A theory building empirical study.* Paper presented at the International System Dynamics Conference, Atlanta, GA.

Moxnes, E. (1998). Not only the tragedy of the commons: Misperceptions of bioeconomics. *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.

Paich, M., & Sterman, J. D. (1993). Boom, Bust, and Failures to Learn in Experimental Markets. *Management Science, 39*(12), 1439-1458.

Piaget, J. (1976). *Die Äquilibration der kognitiven Strukturen*. Stuttgart: Klett.

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

Pirnay-Dummer, P. (2007). *Model Inspection Trace of Concepts and Relations. A Heuristic Approach to Language-Oriented Model Assessment.* Paper presented at the AERA, Division C, TICL SIG, Chicago, IL, USA.

Pirnay-Dummer, P. (2008, 31-03-2008). *Language Oriented Representation of Mental Models.* Paper presented at the Mental Model Workshop, Florida State University, Tallahassee, FL, USA.

Pirnay-Dummer, P. (2010). Complete Structure Comparison. In D. Ifenthaler, P. Pirnay-Dummer & N. M. Seel (Eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge* (pp. 235-258). New York: Springer.

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.

Pirnay-Dummer, P., & Walter, S. (2009). Bridging the World's Knowledge to Individual Knowledge Using Latent Semantic Analysis and Web Ontologies to

Complement Classical and New Knowledge Assessment Technologies. *Technology, Instruction, Cognition and Learning, 7*(1), 21-45.

Richmond, B. (1997). The strategic forum: Aligning objectives, strategy and process. *System Dynamics Review, 13*(2), 131-148.

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.

Schlomske, N., & Pirnay-Dummer, P. (2009). Model based assessment of learning dependent change within a two semester class *Educational Technology Research and Development., 57*(6), 753-765.

Schnotz, W. (1994). *Aufbau von Wissensstrukturen*. Weinheim: Beltz, Psychologie-Verl.-Union.

Schnotz, W., & Preuss, A. (1997). Task-dependent construction of mental models as a basis for conceptual change. Aufgabenabhängige Konstruktion mentaler Modelle als Grundlage konzeptueller Veränderungen. *European Journal of Psychology of Education, 12*(2), 185-211.

Schvaneveldt, R. W., Durso, F. T., Goldsmith, T. E., Breen, T. J., Cooke, N. M., Tucker, R. G., et al. (1985). Measuring the structure of expertise. *International Journal of Man-Maschine Studies, 23*, 699-728.

Seel, N. M. (1991). *Weltwissen und Mentale Modelle*. Göttingen: Hogrefe.

Seel, N. M. (2003). Model centered learning and instruction. *Technology, Instruction, Cognition and Learning, 1(1),* 59-85.

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

Skaza, H., & Stave, K. A. (2009). *A test of the relative effectiveness of systems simulations to increase student understanding of environmental issues.* Paper presented at the 27th International Conference of the System Dynamics Society, Albuquerque, NM.

Skaza, H., & Stave, K. A. (2010). *Assessing the effect of systems simulations on systems understanding in undergraduate environmental science courses.* Paper presented at the 28th International Conference of the System Dynamics Society, Seoul.

Spector, J. M. (2006). Introduction to the special issue on models, simulations and learning in complex domains. *Technology, Instruction, Cognition and Learning, 3*(3-4), 199-204.

Spector, J. M., Christensen, D. L., Sioutine, A. V., & McCormack, D. (2001). Models and simulations for learning in complex domains: using causal loop diagrams for assessment and evaluation. *Computers in Human Behavior, 17*(5-6), 517-545.

Sterman, J. D. (1989). Misperceptions of feedback in dynamic decision making. *Organizational Behavior and Human Decision Processes, 43*(3), 301-335.

Sterman, J. D. (2002). All models are wrong: reflections on becoming a systems scientist. *System Dynamics Review, 18*(4), 501□531.

Sterman, J. D. (2009). *Does formal system dynamics training improve people's understanding of accumulation?* Paper presented at the 27th International Conference of the System Dynamics Society.

Sterman, J. D., & Booth Sweeney, L. (2007). Understanding public complacency about climate change: adults' mental models of climate change violate conservation of matter. *Climatic Change, 80*(3-4), 213-238.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-352.

Appendix

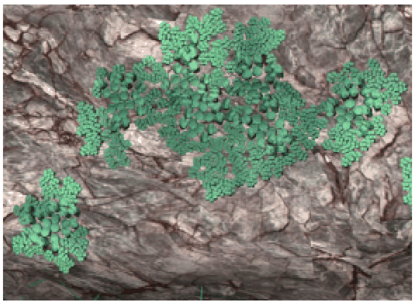
Appendix 1: Instructions for the reindeer rangeland management task (modified with respect to the original instructions in Moxnes, 2004)

For this activity you will play the role of the manager of a reindeer herd. Your task is to produce as many reindeer as possible. But you must also make sure that the animals do not overgraze the lichen, which is the limiting source of food for the reindeer in winter.

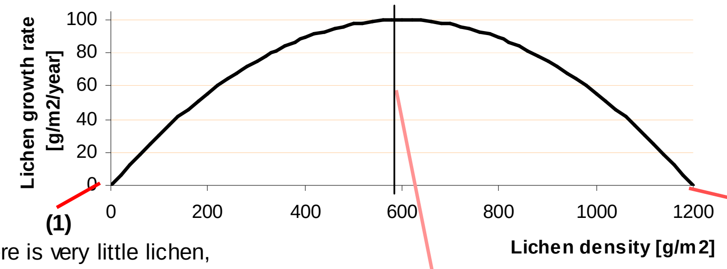
Setting

Your reindeer herd grazes on a pasture used exclusively to feed your herd. Hence its resources will depend only on your decisions regarding the herd size. In summer, food supply is no problem – there is always plenty of grass and herbs. In winter, the food is scare and limited to lichen. If there is no lichen, all the animals will die.

Lichen is a low-growing species that is part plant and part fungus.



Lichen re-grows itself during summer when the reindeer feed on other plants. Lichen grows by propagating its spores. Lichen growth depends on its density and is described by an inverse U-shaped function as illustrated below.



(1)
When there is very little lichen, i.e. when lichen density is very low, there are just a few spores and there will be only little growth.

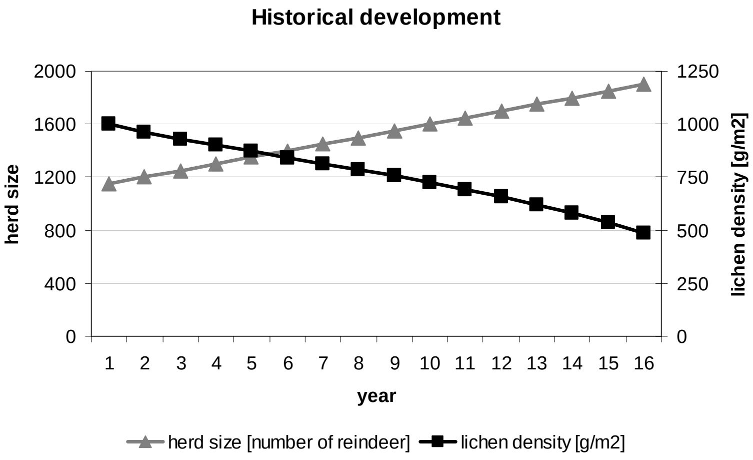
(2)
When there is very much lichen, i.e. when lichen density is very high, they start to fold onto themselves and stop growing.

(3)
In between these extremes, the growth reaches a maximum.

Grazing by reindeer affects lichen density. It therefore also influences the lichen growth rate. You should assume that 1000 reindeer eat 80g/m^2 of lichen during one winter. So as you can see, the reindeer are dependent upon the lichen, but the lichen is dependent upon the reindeer as well. That means that you have to maintain both the reindeer and the lichen populations together.

Starting point

The previous owner has steadily increased the number of reindeer from 1150 to 1900. As a consequence, the lichen density [g/m^2] has dropped from 1000 to 488 g/m^2. This development is shown in the following diagram and table.



Year	lichen density [g/m^2]	herd size [number of reindeer]
1	1000	1150
2	964	1200
3	930	1250
4	900	1300
5	872	1350
6	842	1400
7	814	1450
8	786	1500
9	756	1550
10	726	1600
11	694	1650
12	658	1700
13	622	1750
14	582	1800
15	538	1850
16	488	1900

Decisions to make

It is your job to decide how to maximize the size of your reindeer herd, while maintaining a manageable lichen density. You cannot control the lichen directly. You can control the number of animals you want to keep on the pasture, and that controls the amount of grazing (food eaten) by the animals.

Each year for 15 years, you will set a desired herd size. You are trying to have the maximum number of animals you can, while also maintaining the lichen at the best density for its growth. You should try to achieve the maximum sustainable herd size as soon as possible.

You can vary the herd size freely: You do not have to think about the sex ratio, the number of calves, losses of animals, or the age structure of the herd.

Appendix 2: Instructions for the national development planning task

You have just been elected the Prime Minister of Blendia. You will stay in office as prime minister 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 cost of Africa. It is currently one of the poorest countries in the world with an income per capita of 300 $ per year. Your task is to bring your country onto a sustainable economic growth path and achieve and maintain the highest possible income per capita.

Income per capita results directly from production and production is driven by the available capital (machinery and its technology level) as well as by total factor productivity. As a government you cannot invest in capital directly. However, you can improve the general investment environment. 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. You can specifically invest in the following three resources:

- Education

  Investments in education are used for building and maintaining schooling capacity, i.e., for building and maintaining schools, for training and paying teachers, as well as for paying books.

  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% (maximum or optimal value for Human Development Index calculations).

- Health

  Investments in health are used for building and maintaining basic health care services, i.e., for building and maintaining health care centers, for training and paying doctors and nurses, as well as for paying drugs.

  Health defines the strengths 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 (maximum or optimal value for Human Development Index calculations).

- Roads

  Investments in roads are used for building and maintaining 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 as in the year 2005 in the United States.

Budget issues

For making your investment decisions you will have to take a number of budget mechanisms into account.

Your expenditures for education, health and roads are fed by two sources:

- Revenue: Through taxation the government generates revenue from per capita income.

- Borrowing: You can borrow money from foreign resources. If you borrow money you start accumulating debt. Each year you will have to pay interest on your debt.

Government development expenditure

- In Blendia, government development expenditure is the total revenue minus interest payments on debt.

Decisions

Every five years, as part of a national development planning effort, you decide on the expenditures for education, health and roads. You can do three things and as the prime minister you have the absolute power to decide (see also Figure 1):

1. Distribute more than the total available development expenditure. In this case you borrow money and create a deficit.
2. Distribute less than the total available development expenditure. In this case you will have a surplus and be able to service debt or lend money.
3. Distribute the total available development expenditure without creating neither deficit nor surplus.

Figure 1: Budget decisions mechanism with initial values

Government development expenditure	90 $ per person
– Education expenditure – Health expenditure – Transportation expenditure	– 30 $ per person – 30 $ per person – 30 $ per person
Surplus (+)/deficit (-)	0

Evaluation

Your performance will be evaluated on the following basis:

- Income per capita: You should try to achieve and maintain the highest possible income per capita. The country’s official goal is a value of 600 $ per capita in 50 years from today.
- Interest payments on debt: Per capita income can only be maintained if you have not accumulated excessive debt. At the end of the 50 years period the interest payments on debt in year 50 will be deducted from your income per capita in year 50.

Appendix 3: expert texts for the dynamic decision making tasks

Expert text for the reindeer rangeland management task

“I play the role of the manager of a reindeer herd. I need to produce as many reindeer as possible and I have to reach the maximum sustainable herd size as soon as possible. The limiting source is lichen or lichen density, respectively.

Lichen density increases with lichen growth and decreases with grazing. Grazing depends on the number of reindeer and the grazing per reindeer. Lichen growth depends on lichen density. If lichen density is equal to the optimal lichen density, lichen growth will be equal to maximum lichen growth. If lichen density is above or below the optimal lichen density, lichen growth will be lower than the maximum lichen growth. The relationship between lichen density and lichen growth describes an inverse U-shaped function.

For lichen density to remain stable lichen growth needs to be equal to grazing.

The maximum sustainable herd size can be produced if lichen density is equal to the optimal lichen density and grazing is equal to the maximum lichen growth. The maximum sustainable herd size is 1250 reindeer.

If grazing exceeds lichen growth the number of reindeer has to be reduced; if grazing is below lichen growth the number of reindeer can be increased.

The previous owner has left me an overgrazed pasture with lichen density below optimal lichen density and too many reindeer.”

Detailed and technical expert text for the national development planning task

“I play the role of the prime minister of Blendia, a very poor sub-Saharan African country. My task is to achieve and maintain the highest possible income per capita.

My performance is evaluated by subtracting interest payments on debt from per capita income.

Per capita income is determined by the amount of capital per person and total factor productivity. Capital increases with investment and decreases through depreciation. Investment depends on the potential investment which is the fraction of per capita income used for savings and on the investment environment.

The investment environment improves with higher levels of education, health and roads. Higher levels of education, health and roads also increase total factor productivity.

Education, health and roads improve as a consequence of education, health and roads expenditure. As the prime minister I decide on the desired per capita budget for education, health and roads which together yield the desired per capita development budget.

The available per capita development budget depends on tax revenue and interest payments on debt. Tax revenue is per capital income multiplied by the tax rate.

The difference between the desired per capita development budget and the available per capita development budget determines whether there is a deficit or surplus. In the case of a deficit I need to borrow money and borrowing accumulates debt per capita. In the case of a surplus I can pay back debt per capita.

The higher the debt per capita the more interest has to be paid on debt. Interest payments on debt are subtracted from the tax revenue and thus decrease the available per capita development budget.

Neither roads, health nor education alone can improve the investment environment very much. The investment environment improves fastest when the levels of education, health and roads are similar. Education, health and roads therefore need to develop in a balanced way.

Investments in education take a long time to have an effect. The same holds true for investments in health. The health investment delay is, however, considerably shorter than the education investment delay. The roads investment delay is fairly short.

In order to stimulate a balanced growth of education, health and roads I need to prioritize education in the early years. I also need to invest in roads in the early years because this generates per capita income fairly soon.

If the desired per capita development budget exceeds the available per capita development budget a deficit arises that can only be covered by borrowing. Borrowing adds to debt per capita and leads to exponentially growing interest payments.

At the outset, it is very effective to borrow money and use it to improve education, health and roads. With reasonable debt per capita in the early years and adequate allocation to education, health and roads, per capita income starts growing so well that debt per capita can be paid back and education, health and roads expenditures increase even more.”

Shorter expert text for the national development planning task

Blendia is a very poor country. I have to achieve and maintain the highest possible per capita income.

Per capita income is determined by capital and total factor productivity.

Capital increases with investment and investment increases with higher levels of education, health and roads.

Higher levels of education, health and roads also increase total factor productivity.

Education, health and roads improve as a consequence of education, health and roads expenditure.

I determine the desired expenditure for education, health and roads.

The available expenditure is the tax revenue minus interest payments on debt.

Tax revenue is per capita income multiplied by the tax rate.

The difference between the desired expenditure and the available expenditure determines whether there is a deficit or surplus.

In the case of a deficit I need to borrow money. Borrowing accumulates debt which leads to growing interest payments.

I should balance the levels of education, roads and health because investment increases most with balanced resources.

I should invest early in education because it has the longest delay and therefore takes time to have an effect on per capita income.

I should also invest early in roads because it has a more direct impact on per capita income.

I should also borrow, that is, increase debt, at the beginning because then I have money to invest in education, roads and health right away.

Later I should pay off the debt, after per capita income has improved because interest payments grow exponentially.