

# Seeing with the mind

## The role of spatial ability in inferring dynamic behaviour from graphs and stock and flow diagrams

Guido A. Veldhuis<sup>1\*</sup> and Hubert Korzilius<sup>1</sup>

<sup>1</sup>Institute for Management Research, Radboud University Nijmegen  
P.O. box 9108, 6500 HK Nijmegen, The Netherlands

\* corresponding author: [g.veldhuis@gmx.net](mailto:g.veldhuis@gmx.net), +31623527721

### Abstract

*Several experiments have shown that, when predicting the behaviour of stocks and flows, many participants rely on the erroneous 'correlation heuristic'. They seem to assume that the behaviour pattern of a stock looks similar to that of the flow and vice versa. Based on similar experiments with motion graphs we hypothesize that spatial ability explains variance on tasks involving accumulation. We propose that spatial ability might also generate other important differences between people, such as their ability to infer behaviour from diagrams. We tested participants on two dimensions of spatial ability: visualization and spatial orientation. In an experiment we found that the visualization dimension has a positive effect on performance in various systems thinking inventory tasks and a negative effect on the likelihood that the participant selects a response typical for correlation heuristic reasoning. The positive relation to performance was also present for tasks in which stock behaviour had to be inferred from text and diagrams. Furthermore, we found that people are not persistent in their use of the correlation heuristic between different types of tasks. Males and females did not differ in their spatial ability, but, males did perform better on almost all stock and flow tasks.*

Keywords: Systems thinking inventory task, correlation heuristic, stock-flow failure, spatial ability, stock and flow diagrams, mental simulation, visualization

### Introduction

Throughout the last decade numerous studies have shown that people have difficulties in understanding the most fundamental component of complex systems: accumulation. Starting from the work by Booth Sweeney and Sterman (2000), successive experiments have shown that people have difficulties solving so-called 'systems thinking inventory tasks' (STI tasks). These tasks were designed to test the understanding of accumulation behaviour. The problems people

have with relating the behaviour of stocks and flows were shown in different task designs and appeared not to be attributable to the type of graphs, lack of contextual knowledge, motivation, or cognitive capacity. One of the most common mistakes is the erroneous assumption that the behaviour of the stock matches the pattern of the flows; this reasoning was called the ‘correlation heuristic’ (Cronin et al., 2009).

In the past years more insight in this so-called stock-flow failure was created. Research showed that the term ‘average duration’, which is commonly used in system dynamics (SD), leads to fundamental misunderstanding of continuous versus discrete delays (Grössler et al., 2011; Jacobs et al., 2011). Furthermore, it was found that the reasoning errors that are made are more diverse than just the correlation heuristic and that individual decision making characteristics influence performance (Cronin et al., 2009; Korzilius et al., 2011). Rather encouraging for the field of SD was the finding that a course in SD improved the performance of subjects (Pala and Vennix, 2005; Sterman, 2010) and that even a single lecture might do so (Kainz and Ossimitz, 2002). Research found performance differences related to gender, academic background and origin but has so far not uncovered other underlying factors for explaining differences in performance (Booth Sweeney and Sterman, 2000; Kainz and Ossimitz, 2002; Ossimitz, 2002).

A precise solution strategy for STI tasks is graphical integration and differentiation, although for a simple STI task, such as the so-called department store task, comparing two graph areas is already sufficient to derive the correct solution (Sterman, 2002). Research which has so far not been considered in the SD literature suggests that inferring behaviour from static images, such as diagrams and graphs, places a demand on the visual-spatial processing capacity of people (Hegarty, 2004; Kozhevnikov et al., 2002a, 2002b, 2007; Kozhevnikov and Thornton, 2006; Mayer and Sims, 1994). In this paper we explore to what extent people’s spatial ability is a factor predicting their ability to infer stock and flow behaviour from graphs and diagrams.

## **Spatial ability**

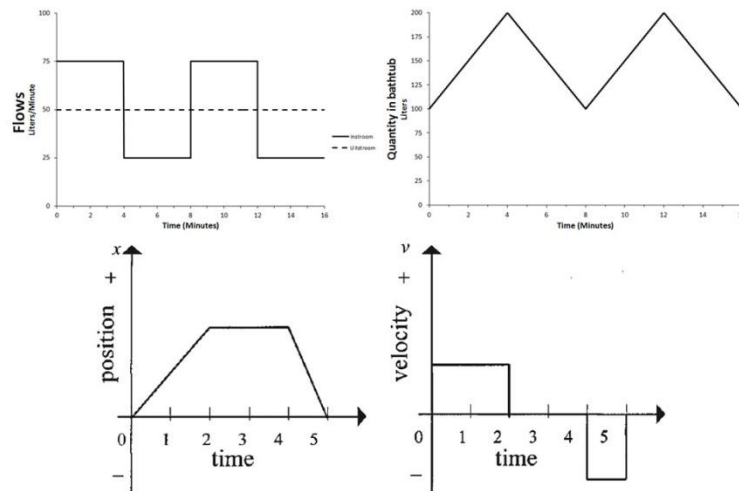
Spatial ability can be defined as the: ‘... *individuals’ abilities in searching the visual field, apprehending the forms, shapes and positions of objects as visually perceived, forming mental representations of those forms, shapes, and positions, and manipulating such representations “mentally”*’ (Carrol, 1993, p. 304). Spatial ability thus goes beyond seeing and apprehending, it involves creating one’s own abstract representations and being able to inspect and change these. This capability is correlated with success in mechanical occupations, mathematics, physics and medical professions (Hegarty, 2004). In a meta-analysis Carrol (1993) identifies five factors that make up spatial ability: Spatial visualization, spatial relations, closure speed, flexibility of closure, and perceptual speed. In this study we include the two relevant factors: visualization and spatial relations.

The factor most often used in studies that include spatial ability measures is visualization. Visualization is: ‘*the ability to manipulate or transform the image of spatial patterns into other arrangements*’ (Ekstrom et al., 1976, p. 173). Visualization tests measure the processes of apprehending, encoding and mentally manipulating spatial forms. The second factor is named spatial relations (Carrol, 1993), sometimes referred to as spatial orientation (Ekstrom et al., 1976, p. 149): ‘*The ability to perceive spatial patterns or to maintain orientation with respect to objects in space.*’ The spatial orientation tests involve rotating objects rather than manipulating them. Research has shown that spatial visualization and

orientation are the dimensions with the highest predictive values for kinematics (motion) graph and physics problems (Kozhevnikov et al., 2002b).

### *Inferring behaviour from graphs*

Based on eye tracker data Carpenter and Shah (1998) propose that a person incrementally forms a larger representation of a graph by looking at unique quantitative relationships (e.g. segments with different slopes). People spend a large amount of time relating information to its graphical referent in cyclical patterns by reading and rereading information from the axes and label regions. It is this process in which people's ability differs and which forms a key factor in explaining why people differ in inferring behaviour from graphs. For a long time research with kinematics tasks (See Figure 1) has provided evidence that a large number of people are less able to translate a position (stock) graph to an abstract representation and subsequently fail to infer the correct velocity (net flow). Instead they view the graph as a literal picture of the event (Barclay, 1986; Beichner, 1994). Later research showed that especially people with low spatial ability make this mistake (Kozhevnikov et al., 2007).



**Figure 1: Tasks revolving around the stock flow relation: the two graphs on the top are a redrawn version of Booth Sweeney and Sterman (2000) 'bath tub' task. The two graphs below are adopted from Kozhevnikov et al. (2002a) 'graph problems' (originally appeared in Beichner, 1994). In the bathtub task participants were presented a graph showing the flows, in the 'graph problem' the behaviour of the stock.**

In repeated experiments, including tasks with the bottom graphs in Figure 1, people with low spatial ability interpreted the graph as a picture: expecting the graph to show exactly what the phenomenon would look like. For example, people would explain a straight segment on a position graph as showing movement forward. Furthermore, this misinterpretation persists when the data displayed (position vs. velocity) changed. People suffering from the 'graph-as-picture' misinterpretation match position graphs with velocity graphs that look similar. The graph-as-picture and correlation heuristic show a striking similarity, both describe the misconception that the behaviour of a stock and a flow should look similar. On the contrary, none of the high spatial ability participants directly referred to the position graph as depicting motion. Subsequent analyses of eye tracking data of these participants suggested that, high spatial ability participants were better able to identify the referent of the plotted information and translate visual patterns into a conceptual relation (Kozhevnikov et al., 2002a, 2007). Similar tasks showed that spatial ability was a predictor for performance on graph tasks before and after an introductory course in physics using computer simulation. In addition, students with high spatial

ability improved more than students with low spatial ability. (Kozhevnikov and Thornton, 2006). Therefore, in this study we hypothesize that:

*H<sub>1</sub>. A person's spatial ability is positively related to the likelihood that their response to a STI task is correct.*

*H<sub>2</sub>. A person's spatial ability is negatively related to the likelihood that their response to a STI task shows a correlation heuristic error.*

### *Inferring behaviour and learning from diagrams*

Diagrams are seen as a way to augment our cognition. This form of external visualization is accompanied by internal visualization, which we might call 'mental simulation'. How well a person is able to perform this activity depends on his or her spatial ability (Hegarty, 2004). To our knowledge no research is conducted on how well people are able to infer behaviour from SD diagrams while at the same time SD diagrams are continuously used to explain exactly those dynamics that we proof people do not understand.

Data from eye-tracking studies reveal that when people look at diagrams and text explaining simple mechanical systems, such as pulley systems, they look at components or groups of connected components and infer the motion of the components one by one and not simultaneously (Hegarty, 1992; Hegarty and Just, 1993; Winn, 1991). Although the context is not analogous to stock flow systems (behaviour is instantaneous, linear and not generated endogenously) these experiments do provide a picture of how dynamic behaviour is inferred from a static diagram. We thus expect that spatial ability predicts how well someone can infer behaviour from diagrams. However, we believe it is unlikely that someone can understand a stock and flow diagram without an explanation.

Most often we explain a system by simultaneously presenting visual information (such as a stock and flow diagram) and verbal information. According to Mayer and Sims (1994) both forms of information are encoded separately in the brain to form mental representations. A third process constructs referential connections that map the structural relations between the two representations of the system. In an empirical study using tasks with simple systems such as a bicycle pump and the human respiration system (note the stocks and flows) it was shown that high spatial ability subjects are able to devote more cognitive resources to making referential connections as opposed to the low spatial ability subject who devote more resources to visual encoding. This leads high spatial ability subjects to profit more from simultaneous presentation of visual and verbal information (Mayer and Sims, 1994). We expect that:

*H<sub>3</sub>. People, who receive a short presentation on stocks and flows, and a stock and flow diagram with each question, perform better than people who do not see a presentation and only receive the questions in text.*

*H<sub>4</sub>. A person's spatial ability is positively related to performance on tasks which use stock and flow diagrams after these concepts have been explained in a presentation.*

Most past studies only included one task, those who did include multiple tasks often did not report on the relationships between responses on different tasks. Since we include multiple tasks we have formulated the following *additional* research questions.

*ARQ<sub>1</sub>. What is the relation between correct responses on different tasks?*

*ARQ<sub>2</sub>. What is the relation between correlation heuristic responses on different tasks?*

As mentioned in the introduction, gender had a predictive value in some past studies. However, results vary and are often not reported. To help create a more complete picture of the role of gender we ask the following additional research questions:

- ARQ<sub>3</sub>. What is the effect of gender on the likelihood that a correct response is provided?*  
*ARQ<sub>4</sub>. What is the effect of gender on the likelihood that a correlation heuristic response is provided?*

## Method

### Participants

The participants ( $N = 88$ ) for this study were recruited by e-mail within our personal network and through an undergraduate course in statistics at the Management science faculty of the Radboud university. Participants were asked to follow a link to an online survey that would require 30 minutes of their time. The content of the survey was described as a number of ‘fun’ and ‘challenging’ tasks. From the complete responses, we have drawn two winners of two 20 euro gift certificates.

People who reported having studied system dynamics concepts or who reported recognizing one of the tasks were removed from the study. Furthermore, we removed participants that: were outliers in age ( $< 18, 30 >$  years), answered a task extremely fast ( $< 20$  seconds), skipped the video explanation or who reported having made part of the experiment twice due to technical errors. Our sample consists of highly educated young adults in which female outnumbered male participants. An overview of the demographic characteristics of the sample is provided in Table 1.

**Table 1. Participant demographics ( $N=88$ )**

	<i>n</i>	<i>%</i>		<i>n</i>	<i>%</i>
<b>Age</b>			<b>Gender</b>		
18-21	11	12.5	Male	39	44.3
21-24	31	35.2	Female	49	55.7
24-27	43	48.9			
27-30	3	3.4			
<b>Highest obtained degree</b>			<b>Field of study<sup>1</sup></b>		
High school	6	6.8	Business	30	34.1
Vocational education	2	2.3	Social Sciences	18	20.5
University of applied science	7	8	Economics	11	12.5
Research university			<b>Condition</b>		
Propaedeutic	14	15.9	Control	49	55.7
Bachelor degree	21	23.9	Manipulation	39	44.3
Master degree	38	43.2			

**Note:** <sup>1</sup> Only the three largest groups are reported

### Procedure

The research design is of a psychometric nature. To conduct our experiment we developed an online survey using the Qualtrics web based tool<sup>1</sup>. This did mean transforming

<sup>1</sup> Qualtrics Labs inc. software, Version 27244 of the Qualtrics research suite, copyright 2011 Qualtrics labs inc. [www.qualtrics.com](http://www.qualtrics.com)

some traditional pencil and paper test to a digital format. We will discuss adjustments made for each test. Furthermore, our research was conducted in the Netherlands, and our wish was to make the tasks as easily understandable as possible, this lead us to translate all pre-existing material into Dutch. The experiment is composed of three separate parts: 1. Measurements for spatial ability, 2. Systems thinking inventory tasks, 3. Inferring behaviour from text and diagrams. Part one, which contains tests for spatial ability, was time restricted and used an auto-advance option. Participants could the tasks in second and third part at their own pace; they were shown one task at a time and could proceed to the next task by clicking a forward button. Once participants had finished a question they could not go back to alter their response.

In the third part of the experiment participants were automatically assigned to a control or manipulation condition. In the manipulation condition participants were first shown a 4-minute narrated slideshow presentation. The presentation explained stocks, (constant) flows and feedback and their diagrammatic notationality through stories about a bathtub and a mice population. After this the participants were presented with the final two tasks, in addition to a written description, participants also received a stock and flow diagram (Figure 4). The tasks combine the elements creating behaviour (stable inflow, balancing and reinforcing loops) in a different way than shown in the presentation. Participants in the control condition were not shown the presentation or the diagrams.

### Materials

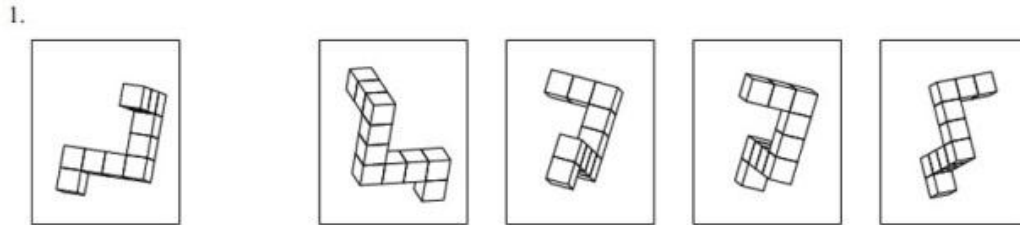
*Spatial ability.* We included two tests for spatial ability: the paper folding test (PFT) and the mental rotations test (MRT). Carrol (1993) identifies the PFT as one of the most widely used tests for visualization, it is often used in relevant research, as a composite with other measures, or as the sole measurement of spatial ability (Kozhevnikov et al., 2002a; Kozhevnikov and Thornton, 2006; Kozhevnikov et al., 2007). The PFT shows the participant how a piece of paper is folded a number of times and how one or multiple holes are punched in the folded paper. The participant should subsequently pick a picture of what the paper would look like unfolded from 5 alternatives. The test has two parts, of 10 items each; there is a 3 minute time limit for both (Ekstrom et al., 1976). Occasionally, such as in the research cited above, just one part of the test is used and yields reliable results. We used part I of the original test. People were first presented a screen with the content of the first sheet of paper containing explanation and example problems. Figure 2 shows one of the example problems. The test screen contained all 10 problems which is similar to the pencil and paper test. Each participants score was calculated by summing the number of correct answers. The PFT contains items of varying difficulty, the simplest items generated no or low variance and differences mainly originated in differing speed at which they could answer the questions. This resulted in a rather low Cronbach's  $\alpha$  of .63 (see Table 2), we will discuss the reliability of this scale in more depth in the discussion.



**Figure 2. Paper Folding Task Example: adopted from (Ekstrom et al., 1976, p. 176).**

We included a measurement of the spatial relations dimension using the Vandenberg and Kuse (1978) mental rotations test (MRT). The MRT is based on the two dimensional drawings of three dimensional objects originally developed by Shephard and Metzler (1971). The original test contained 20 items. Each item consists of one target object, two correct alternatives and two 'distractors'. Participants have to select both of the two correct alternatives, which show the same object as the target figure, rotated, so that it is shown from a different perspective.

Participants were first shown a briefing screen that prompted them to try mentally rotating a figure and to attempt three practise questions. On the next screen the first 10 items of the redrawn test (Peters et al., 1995) were administered. An example question is shown in Figure 3. The score per participants was calculated by summing the number of questions in which both correct alternatives were marked. The test resulted in a Cronbach's  $\alpha$  of .79, indicating a moderately high reliability.



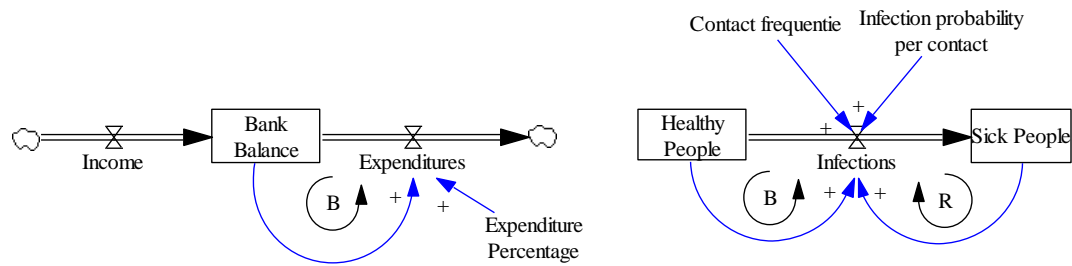
**Figure 3. Example question of the Redrawn MRT adopted from Peters et al. (1995).**

*Systems thinking inventory tasks.* We included three STI tasks. To be able to shed new light on the performance on two well-known tasks we first included the department store task (Serman, 2002). We will refer to the other two tasks of the first part as ‘graph tasks’ since participants are shown a graph and subsequently have to select the correct associated graph. All graphs can be found in Appendix A.

The first graph task is the ‘square wave’ bath tub task 1 (Booth Sweeney and Serman, 2000). We provided a redrawn version of the task as shown in Figure 1. To make the online format possible we prepared six multiple choice options. We based these six options on examples of responses provided by Booth Sweeney and Serman (2000) and expected common mistakes.

After the bath tub tasks we added a new cash flow task based on Beichner’s (1994) task 11. In this task participants should choose the appropriate graph showing the cash flow (net flow) for a given bank balance (stock) graph. This task allows us to reproduce the kinematics experiment in a context more common for SD. The three STI tasks provided us with two outcome measures each. First, if the participants answered the question correct or incorrect and second, if the participants selected a response typical for correlation heuristic reasoning.

*Inferring behaviour from text and diagrams.* This part of the experiment included two tasks. Participants were given a description of the problem situation by text or by text and a diagram (Figure 4). The first task involves a bank account with a stable inflow and balancing outflow creating goal seeking behaviour. The second task is based on the SIR model as discussed by Serman (2000). We rephrased some variables to make the task easier to understand (e.g. healthy instead of susceptible). This model shows the spread of an infectious disease in an S-shaped pattern. The outcome measure of both tasks was incorrect/correct. A detailed overview of the tasks can be found in Appendix B.



**Figure 4** Diagrams of the ‘bank balance’ task and the ‘epidemic task’, the latter is adapted from Sterman (2000).

### Data Analyses

We conducted independent sample t-tests when comparing means and Chi square tests when comparing categorical variables; in the latter case we used Cramer’s V to measure the strength of the association. When testing the predictive value of multiple variables for a categorical outcome we use forced entry logistic regression. This method provides us the opportunity to use categorical (gender) and ratio level (PFT score) predictors to determine the odds of a participant selecting a categorical outcome. However, while in the STI task part the sample size achieved a satisfactory number of ‘events per variable’ (EPV) it did not when we further divided the sample in conditions. With an EPV of less than 10 the risk of over or underestimating the regression coefficient increases (Peduzzi et al., 1996). Therefore we do not perform the logistic regression analyses for the separate conditions in the third part of the experiment. Also performing t-tests for subcategories, for example a test of the PFT score for males vs. females for correct answers in one of the conditions would result in low statistical power.

## Results

### Descriptives

Descriptive results for all tasks can be found in Table 2. For the PFT there were no differences between males ( $M=6.72$ ,  $SD=1.93$ ) and females ( $M=6.88$ ,  $SD=1.71$ ),  $t(86)=0.41$ ,  $p>.1$ , or the MRT ( $M_{males}=6.56$ ,  $SD_{males}=2.44$ ;  $M_{females}=5.86$ ,  $SD_{females}=2.52$ ),  $t(86)=1.33$ ,  $p>.1$ . The MRT and PFT measure closely related dimension and as a result are strongly correlated,  $r=.43$ ,  $p<.001$ .

### Systems thinking inventory tasks

*Department store task.* There was a gender effect for department store task Q3,  $\chi^2(1)=11.31$ ,  $p<.01$ , with males outperforming females. This effect is also present for Q4 although weaker and only marginally significant,  $\chi^2(1)=3.79$ ,  $p<.1$ . Considerably more women than men gave a correlation heuristic response on department store Q3,  $\chi^2(1)=10.91$ ,  $p<.001$ ) and department store Q4,  $\chi^2(1)=10.23$ ,  $p<.001$ . We did not find a relation between the measures for spatial ability and correct/incorrect or correlation heuristic responses. The results of the logistic regression analyses can be found in Table 3, this table also reports which answers are correct and which show correlation heuristic reasoning.

*Bathtub task.* A detailed overview of the results of all graph tasks can be found in Appendix A. In the bathtub task more male than female participants selected the correct alternative, although this effect was only marginally significant  $\chi^2(1)=3.03$ ,  $p<.1$ . The logistic



regression showed that the odds of selecting the correct answer increased as the PFT score increased. The higher the PFT score the lower the odds for a correlation heuristic error (answers D and F).

**Table 2. Descriptive results of all tasks**

<b>Spatial ability task results</b>	<i>M</i>	<i>Min-Max</i>	<i>SD</i>	Reliability ( $\alpha$ )
Paper Folding Task	6.81	2 – 10	1.81	.63
Mental Rotation Task	6.17	0 – 10	2.49	.79

<b>System inventory task results</b>	Correct		Correlation heuristic	
	<i>n</i>	%	<i>n</i>	%
Department store Q3	29	33	30	34.1
Department store Q4	17	19.3	24	27.3
Bathtub task	45	51.1	17	19.3
Cash flow task	47	53.4	32	36.3

<b>Behaviour from text and diagrams</b>	Correct			
	Manipulation		Control	
	<i>n</i>	%	<i>n</i>	%
Bank balance task	16	41	20	40.8
Epidemic task	13	33.3	13	26.5

*Cash flow task.* In the cash flow task we again see a strong gender effect favouring males,  $\chi^2(1) = 9.52$ ,  $p < .01$ . The logistic regression furthermore shows that the higher a participant's PFT score the greater the odds are that he or she will select a correct response. Both gender and PFT score predict whether or not someone selects a correlation heuristic response. These results indicate that, again, the odds are higher that a female provides a correlation heuristic as well as people who score lower on the PFT.

Performance between all of the STI tasks is correlated (see Table 4) Overall only 7 people (8%) answered all 4 questions correctly, the average was 1.57 ( $SD = 1.28$ ). Table 5 shows the correlation matrix for correlation heuristic responses. The use of the correlation heuristic is significant within the same type of task (e.g. department store) but not between different types of tasks, the correlation between the two department store tasks is strong ( $V = .574$ ,  $p < .001$ ) but between the two graph tasks rather weak ( $V = .228$ ,  $p < .05$ ).

**Table 3. Overview of the forced entry logistic regression on Incorrect, Correct and Correlation Heuristic (No/Yes) responses on the STI tasks**

	DepStoreQ3			DepStoreQ4			Bathtub Task			Cash flow Task		
	<i>B</i>	<i>SE β</i>	<i>Odds Ratio</i>	<i>β</i>	<i>SE β</i>	<i>Odds Ratio</i>	<i>β</i>	<i>SE β</i>	<i>Odds Ratio</i>	<i>β</i>	<i>SE β</i>	<i>Odds Ratio</i>
<b>Outcome: Correct Answer</b>	T = 13			T = 30			C			B		
<b>Predictors</b>												
Gender <sup>1</sup>	1.65**	0.50	5.18	1.07 <sup>†</sup>	0.57	2.93	0.96*	0.16	2.61	1.65**	0.51	5.19
PFT Score	0.10	0.16	1.11	0.01	0.17	1.01	0.45**	0.16	1.58	0.40*	0.16	1.50
MRT Score	-0.04	0.12	0.96	0.00	0.13	1.00	-0.04	0.11	0.96	-0.06	0.11	0.94
<i>R</i> <sup>2</sup> , Cox & Snell, Nagelkerke	.13, .18			.04, .07			.15, .19			.18, .24		
<b>Outcome: Correlation Heuristic Answer</b>	T = 8			T = 17			D, F			A, C		
<b>Predictors</b>												
Gender <sup>1</sup>	-1.62**	0.53	0.20	-1.78**	0.61	0.17	-0.79	0.62	0.46	-1.00*	0.50	0.37
PFT Score	0.02	0.15	1.02	0.05	0.16	1.05	-0.51**	0.19	0.60	-0.31*	0.15	0.73
MRT Score	-0.07	0.11	0.93	-0.02	-0.02	0.98	0.17	0.14	1.18	0.04	0.11	1.04
<i>R</i> <sup>2</sup> , Cox & Snell, Nagelkerke	.13, .18			.12, .17			.09, .15			.09, .13		

**Note:** <sup>1</sup> 0 = female, 1 = male, <sup>†</sup>*p* < .1, \**p* < .05, \*\**p* < .0

**Table 4. Correlations between incorrect/correct responses on the STI tasks**

Variables	1	2	3	4	5	6
1. Department Store Q3	–					
2. Department Store Q4	.57***	–				
3. Bathtub Task	.29**	.13	–			
4. Cash flow Task	.31**	.11	.27*	–		
5. Bank Balance Task	.37**	.31**	.31**	.27*	–	
6. Epidemic Task	-.09	-.13	-.11	.06	.12	–

Note: \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$

**Table 5. Correlations between responses showing the correlation heuristic reasoning error (no/yes) on the STI tasks**

Variables	1	2	3	4
1. Department Store Q3	–			
2. Department Store Q4	.74***	–		
3. Bathtub task	.07	-.04	–	
4. Cash net flow Task	.15	.12	.23*	–

Note: \* $p < .05$ , \*\*\* $p < .001$

#### *Behaviour from text and diagrams*

For this part of the experiment participants were randomly assigned to either the manipulation condition or the control group, detailed results can be found in Appendix B.

*Bank balance task.* Overall there was no significant difference in performance between the conditions in the bank balance task. However, males in the manipulation condition performed significantly better than females,  $\chi^2(1) = 3.95$ ,  $p < .05$ , as well as compared to men in the control condition (58.8% correct vs. 40.9% correct), although the latter effect was non-significant. Furthermore, within the manipulation condition, those who solved the task correctly had a significantly higher PFT scores ( $M = 7.37$ ,  $SD = 1.89$ ) than those who gave an incorrect response ( $M = 6.09$ ,  $SD = 1.76$ ),  $t(37) = 2.18$ ,  $p < .05$ . In the control condition the difference in PFT score between those that answered correctly ( $M = 7.45$ ,  $SD = 1.23$ ) and incorrectly ( $M = 6.6$ ,  $SD = 1.95$ ) was smaller and only marginally significant,  $t(48) = 1.68$ ,  $p < .1$ .

*Epidemic task.* We also did not find a significant difference in performance between the two conditions in the epidemic task. For male participants performance was higher in the manipulation condition (35.3% correct) compared to the control condition (18.2% correct), this effect however was non-significant. Although non-significant, women performed better than men in the control condition (33.3% vs. 18.2% correct). Overall performance was lowest of all graph tasks with only 29.5% of the participants answering the question correctly. An interesting result was that 50% of the participants selected option F which shows only exponential behaviour.

## **Discussion**

In this study we found that visualization, a dimension of spatial ability, predicts performance in various tasks revolving around the accumulation process. On the contrary, all results seem to indicate that the spatial relations dimension is irrelevant, this partially confirms

hypothesis 1. The more able a person is at visualizing the better he or she performs at STI tasks, related to this finding was the result that low visualization ability increases the odds that a person provides an incorrect answers showing correlation heuristic reasoning. This partially confirms hypothesis 2 and is in line with previous research (Kozhevnikov et al., 2007). Our results indicate that visualization determines how well someone is able to translate visual patterns into conceptual relations. Recently studies in the field of SD have used think-aloud protocols, future studies could focus on determining how visualization ability influences reasoning. These findings could be supported by eye tracker data to determine if people with different visualization abilities indeed look at graphs differently. When people look at graphs differently they might benefit from different graph layouts or oral feedback.

We expected that people with high spatial ability are better at inferring behaviour from diagrams and learning from presentation aided by stock and flow diagrams. The results of the epidemic task are inconclusive, this task is particularly difficult, participants get confused and the majority chose an exponential only answer. Results from the bank balance task showed that a short explanation about stocks and flows and the support of diagrams had little effect on people their understanding of the task, this rejects Hypothesis 3. However, there was some weak evidence that males did benefit from the manipulation. Performance in the manipulation condition was tied stronger to visualization than in the control condition, this provides some indication that participants with high visualization ability are able to benefit more from the manipulation. These findings from the bank task provide only weak support for hypothesis 4, again the spatial relations dimensions appears irrelevant. Future research could focus on how well people understand systems explained in a presentation. Furthermore, eye tracker data could again be useful to determine how people of varying ability and experience look at stock and flow diagrams. An important question to ask would be if experience and education in SD offsets the differences between individuals originating from visualization and other abilities, or if the differences in performance persist.

In order to assess how the results of multiple tasks relate to each other we formulated two additional research questions. The first question investigates how overall performance is related between different tasks. Performance between most of the STI tasks is correlated (see table 4), this indicates that people their performance between tasks does show evidence of an underlying degree of ability but only to a certain extent; the relative weak strength of the correlations indicates that answering one of the tasks (in)correct is no guarantee for performance on another task.

Previous research demonstrated that the correlation heuristic appears in a wide variety of task designs (Cronin et al., 2009), however, little is known about if participants use the correlation heuristic repeatedly from task to task. This lead us to pose an additional research question. The results show that people do not use the correlation heuristic persistently from one task to the next, this holds especially true when tasks are different (see table 5). We found a strong relationship between correlation heuristic responses within the department store task and a modest relation between the two graph tasks. We did not find such a relation between the graph tasks and the department store task. Future research could determine if people do use the heuristic persistently when solving similar tasks. The multiple choice format used in this research can help make the process of collecting and analysing data more practical. Although hand drawn responses provide a richer picture of the participants ability, our format seems comparable since it provides results that fall in to the range of previous paper-pencil tasks and also delivers the expected erroneous responses<sup>2</sup>.

---

<sup>2</sup> For an overview of some department store task results see: Pala and Vennix (2005), for various bathtub task results see: Capelo and Dias (2005) or Kapmeier (2004).

The final two research questions focused on gender differences. We found a very strong gender effect in almost all tasks, both on the incorrect/correct outcome measure and the correlation heuristic measure. Female participants more often provided an incorrect and correlation heuristic response. Gender differences have been found to varying degrees by other studies (Booth Sweeney & Sterman, 2000; Kainz and Ossimitz, 2002; Ossimitz, 2002). These results have two major implications. First, analyses of STI tasks results should be done in the light of potential gender effect. Second, more research is necessary to pinpoint the cause of the gender effect (our research provides no evidence of spatial ability being related to these differences). The latter is especially important since the STI tasks are very similar to current methods used to explain stock/flow behaviour and test student their understanding of it. These methods might severely favour male subjects.

Our findings should be viewed in the light of some limitations. The relative low alpha of the PFT is cause for concern. People can use analytic strategies and mental visualization strategies to solve visualization tests like the PFT. Tests that are speeded, difficult and present items simultaneous favour visual strategies over analytic ones. People who do use analytic strategies switch to visual strategies when item difficulty in the PFT test increases (Kyllonen et al., 1984; Ullstadius et al., 2004). To determine if the PFT indeed tested participants visualization ability we have run all logistic models again with only the 5 most difficult PFT items and found that this only increased the effect sizes while staying significant for almost all outcome variables. This, together with the fact that it is a validated and often used scale, bolsters our trust in the validity of the visualization measurement. Future research should use more elaborate measures of visualization, while keeping in mind other aptitudes and cognitive preferences. This will allow the formation a more encompassing view of what determines performance on stock flow tasks. For example, the broader PPIK model already showed an influence of intelligence and knowledge on inventory management performance (Strohhecker & Grossler, 2011). Individual preferences for a problem solving style might also play a crucial role; a dichotomy of participants in verbalizers and visualizers has provided result indicating that the tendency for low spatial ability subjects to choose correlation heuristic responses only applies for those following visual strategies (Kozhevnikov et al., 2002a).

## **Conclusion**

This study showed that the ‘mind’s eye’ plays an important role in recognizing visual patterns in graphs and diagrams and translating these patterns to abstract concepts in order to infer correct dynamic behaviour. We furthermore showed that participants do not use the erroneous correlation heuristic persistently from task to task. Furthermore, females seem to be more inclined to rely on this erroneous heuristic and perform worse overall. Future studies can more precisely determine what the effect is of different abilities, attitudes and strategies on inferring dynamic behaviour from graphs and diagrams. This can help us improve the way we teach system dynamics and present our results. It can help us remove potential biases favouring a gender or a group of people with different abilities. This paper has contributed to the growing body of literature by exploring a new factor: spatial ability.

## **References**

- Barclay, WL. 1986. Graphing misconceptions and possible remedies using microcomputer-based labs. In *Proceedings of the 1986 National Educational Computing Conference*. San Diego.

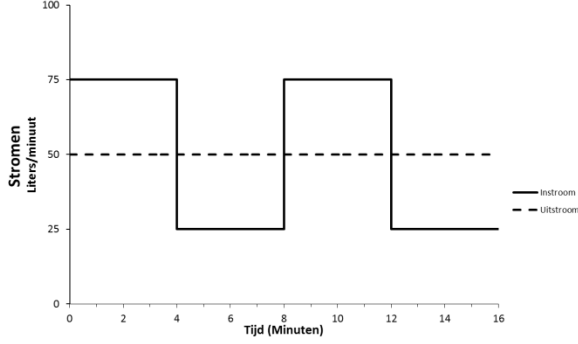
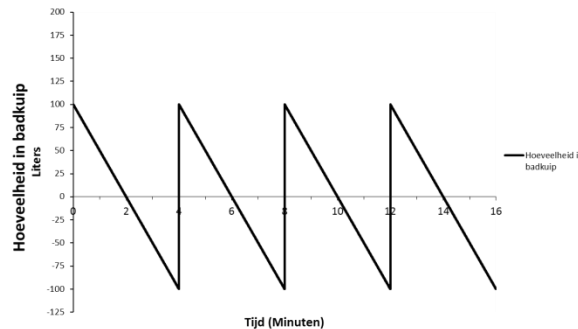
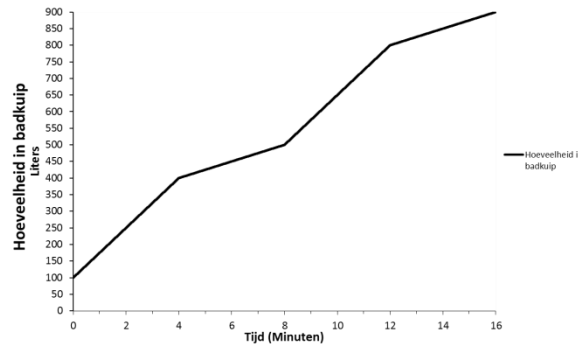
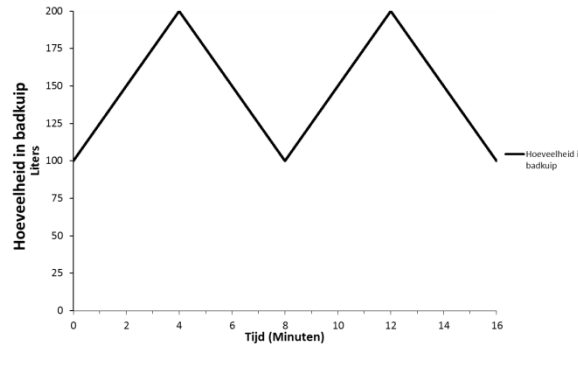
- Beichner, RJ. 1994. Testing student interpretation of kinematics graphs. *American Association of Physics Teachers* **62**(8): 750-762.
- Booth Sweeney L, Sterman JD. 2000. Bathtub dynamics: initial results of a systems thinking inventory. *System Dynamics Review* **16**: 249-286.
- Capelo C, Dias, J. 2005. Strategy lab experiences: bathtub dynamics. In *Proceedings of the International Conference on Advances in Management*. Washington D.C.
- Carpenter P, Shah P. 1998. A model of the perceptual and conceptual processes in graph Comprehension. *Journal of Experimental Psychology* **42**: 75-100.
- Carroll J. 1993. *Human Cognitive Abilities: a survey of factor-analytical studies*. Cambridge University Press: New York.
- Cronin MA, Gonzalez C, Sterman JD. 2009. Why don't well-educated adults understand accumulation? A challenge to researchers, educators, and citizens. *Organizational Behaviour and Human Decision Processes* **108**: 116-130.
- Ekstrom R, French J, Harman H, Dermen D. 1976. *Manual for Kit of Factor-Referenced Cognitive Tests*. Educational Testing Service: Princeton.
- Grössler A, Bleijenbergh I, Vennix J. 2011. "10 years on average doesn't mean 10 years in any case" - an experimental investigation of people's understanding of fixed and continuous delays. In *Proceedings of the 29th International Conference of the System Dynamics Society*, Washington.
- Hegarty M. 1992. Mental animation: inferring motion from static displays of mechanical systems. *Journal of Experimental Psychology* **185**: 1084-1102.
- Hegarty M. 2004. Diagrams in the mind and in the world: Relations between internal and external visualizations. In *Diagrammatic representation and inference*, Blackwell A, Mariott K, Shimojima A. Springer-Verlag: Berlin; pp. 1-13.
- Hegarty M, Just M. 1993. Constructing mental models of machines from text and diagrams. *Journal of Memory and Language* **32**: 717-742.
- Hegarty M, Kozhevnikov M. 1999. Types of visual-spatial representations and mathematical problem solving. *Journal of Educational Psychology* **91**(4): 684-689.
- Jacobs D, Bleijenbergh IL, Vennix JAM. 2011. Supporting decision-makers in managing stock-flow problems: The effects of oral feedback on reasoning. In *Proceedings of the 29th International Conference of the System Dynamic Society*, Washington DC.
- Kainz D, Ossimitz G. 2002. Can Students Learn Stock-Flow-Thinking? An empirical investigation. In *Proceedings of the 20th International Conference of the System Dynamics Society*. Palermo.
- Kapmeier F. 2004. Findings from four years of bathtub dynamics at a higher management education institutions in Stuttgart. In *Proceedings of the 22th International System Dynamics Conference*. Oxford.
- Korzilius H, Raaijmakers S, Rouwette E, Vennix J. 2011. In search of explanation for stock-flow performance. In *Proceedings of the 29th International Conference of the System Dynamics Society*: Washington DC.
- Kozhevnikov M, Thornton R. 2006. Real-Time data display spatial visualization ability and learning force and motion concepts. *Journal of Science Education and Technology* **15**(1): 111-132.
- Kozhevnikov M, Hegarty M, Mayer R. 2002a. Revising the visualizer-verbalizer dimension: evidence for two types of visualizers. *Cognition and Instruction* **21**(1): 47-77.

- Kozhevnikov M, Hegarty M, Mayer R. 2002b. Visual/Spatial abilities in problem solving in physics. In *Diagrammatic Representation and Reasoning*, Anderson M, Meyer, B, Olivier, P. Springer-Verlag: New York: 155-173.
- Kozhevnikov M, Motes M, Hegarty M. 2007. Spatial visualization in physics problem solving. *Cognitive science: A Multidisciplinary Journal* **31**(4): 549-579.
- Kyllonen P, Lohman D, Snow R. 1984. Effects of aptitudes, strategy training, and task facets on spatial task performance. *Journal of Educational Psychology* **76**(1): 130-145.
- Mayer R, Sims V. 1994. For whom is a picture worth a thousand words? Extensions of a dual coding theory of multimedia learning. *Journal of Educational Psychology* **86**(3): 389-401.
- Ossimit G. 2002. Stock-flow-thinking and reading stock-flow-related graphs: An empirical investigation in dynamic thinking abilities. In *Proceeding of the 20th International Conference of the System Dynamic Society*. Palermo, Italy.
- Pala O, Vennix J. 2005. Effect of system dynamics education on system thinking inventory task performance. *System Dynamics Review* **21**(2): 147-172.
- Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. 1996. A simulation study of the number of events per variable in logistic regression analyses. *Journal of Clinical Epidemiology* **49**(12): 1373-1379.
- Peters M, Laeng B, Latham K, Jackson MZ, Richardson C. 1995. A Redrawn Vandenberg and Kuse Mental Rotations test: Different versions and factors that affect performance. *Brain and Cognition* **28**: 39-58.
- Shepard R, Metzler J. 1971. Mental rotation of three-dimensional objects. *Science* **171**(3972): 701-703.
- Sterman J. 2000. *Business Dynamics: System thinking and modeling for a complex world*. McGraw-Hill: Boston.
- Sterman J. 2002. All models are wrong: reflections on becoming a systems scientist. *System Dynamics Review* **18**(4): 501-531.
- Sterman J. 2010. Does formal system dynamics training improve people's understanding of accumulation? *System Dynamics Review* **26**(4): 316-334.
- Strohhecker J. 2009. Does a better understanding of accumulation indeed predict a higher performance in stock flow management? In *Proceedings of the 27th International Conference of the System Dynamics Society*, Albuquerque.
- Strohhecker J, Grossler A. 2011. Intelligence, personality, and interests – Determinants of individual inventory management performance? In *Proceedings of the 18th EurOMA conference*. Cambridge.
- Ullstadius E, Carlstedt B, Gustafsson J. 2004. Multidimensional item analysis of ability factors in spatial test items. *Personality and individual differences* **37**: 1003-1012.
- Vandenberg S, Kuse A. 1978. Mental rotations a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills* **47**: 599-604.
- Winn WD. 1991. Learning from maps and diagrams. *Educational Psychology Review* **3**(3): 211-247.

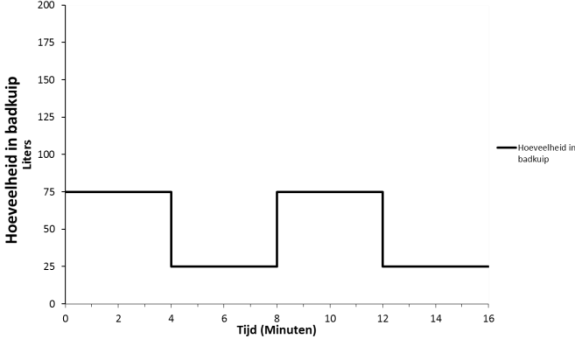
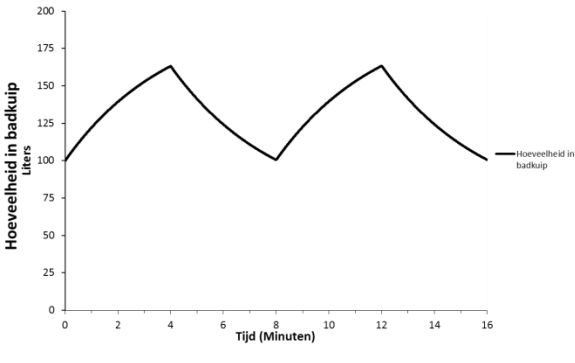
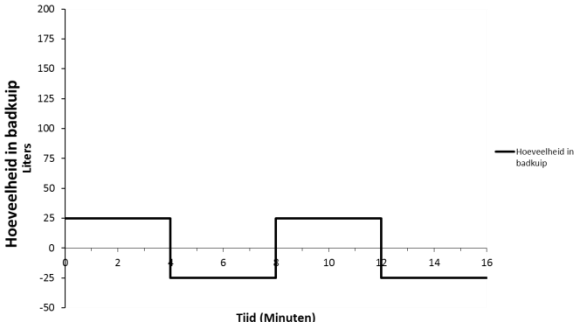
## Appendix A – Systems thinking inventory tasks

### Detailed results Bath tub task

The bath tub task is adopted from Booth Sweeny and Sterman (2000), for this experiment a multiple choice format was developed. The question was translated to Dutch and the question graph was redrawn.


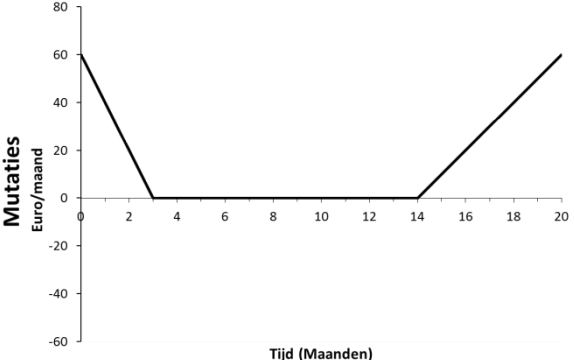
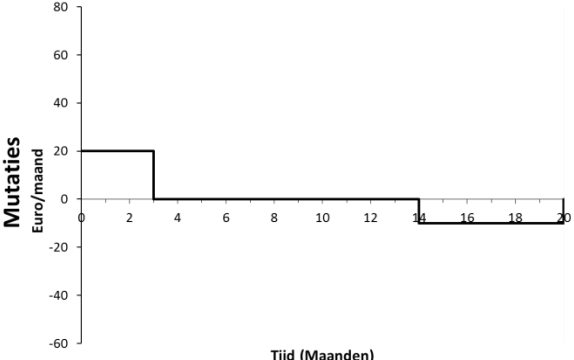
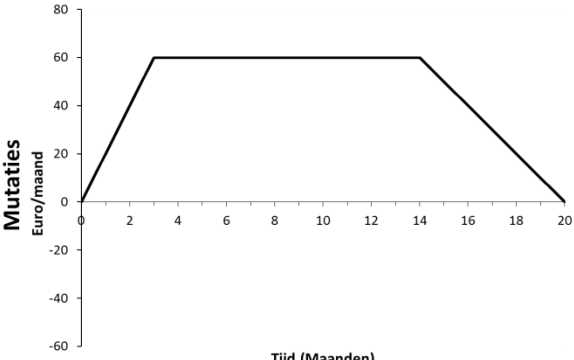
Question		Participants were asked to select one of six alternative graphs that showed the amount of water in the bathtub. Their choice should be based on the information provided in the graph (showing the inflow and outflow) and an initial condition of 100 litres.	
A		Count (%)	Pattern
		4 (4%)	A constant outflow of 50 litres/minute and a pulse inflow of 100 litres every 4 minutes. The outflow is correct but the inflow shows no resemblance to the information provided.
B		7 (8%)	This graph shows the result of a correct inflow but no outflow. This leads to an ever growing amount of water in the bathtub, at various (correct) rates.
C		45 (51%)	The correct answer. The stock rises linearly when the inflow exceeds the outflow and falls in the same fashion when outflow exceeds inflow. Due to the symmetry in the area covered by the net flow the peaks and valleys occur every 4 minutes and are always at 100 and 200 litres.

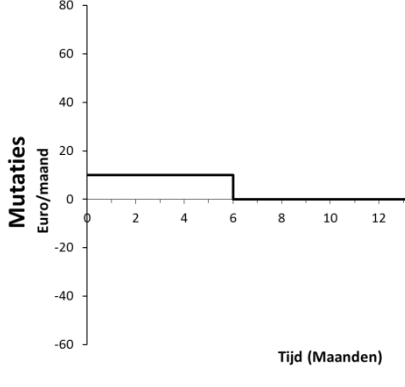
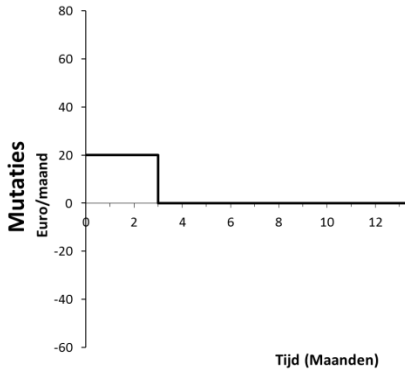
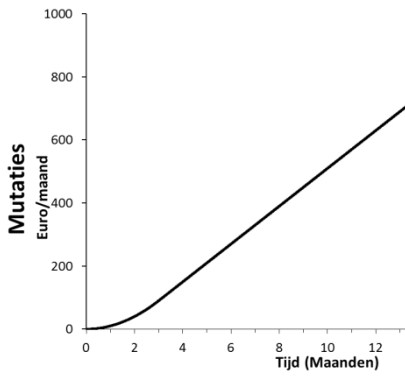


D		6 (7%)	A graph showing correlation heuristic reasoning. This graph is a copy of the inflow graph. It misperceives the relationship between a stock and its flows, resulting in identical behaviour while completely ignoring the outflow.
E		15 (17%)	This graph switches back and forth between goal seeking behaviour towards 200 and 100 litres, switching every 4 minutes. This results in correct peaks and valleys. However, it does not show the correct relation between a constant flow and the appropriate stock behaviour, resulting in incorrect slopes.
F		11 (13%)	A correlation heuristic response. It shows the net flow behaviour, e.g. inflow - outflow. Shows discontinuous behaviour for the stock, wrong slopes and incorrect heights of peaks and valleys. Considers the stock behaviour and flow behaviour to be similar.

### Detailed results of the cash flow task

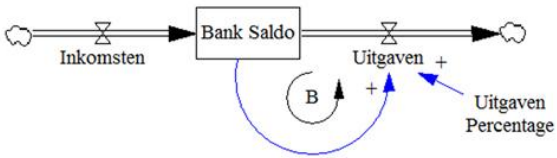
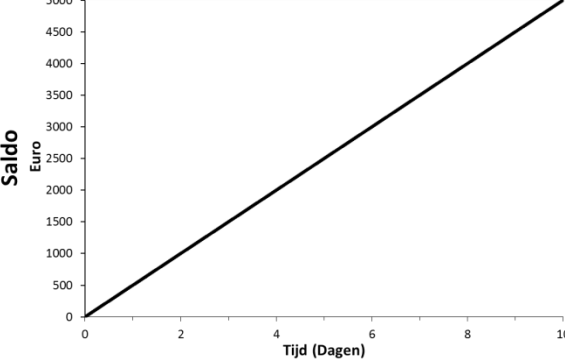
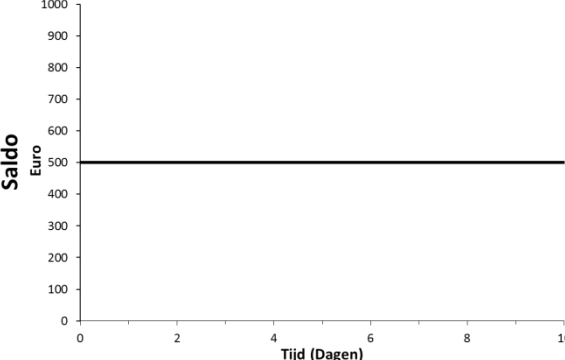
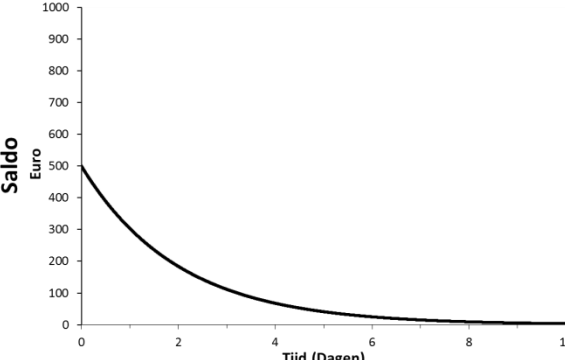
The cash flow task is based on task 11 from Beichner (1994). It has been used both in a multiple choice format and in a think aloud setting (Kozhevnikov et al., 2002a, 2007). We added some options, changed the language to Dutch, redrawn the graphs and added numerical values to the axis.

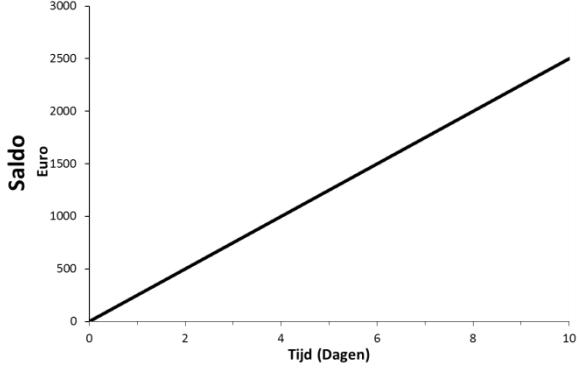
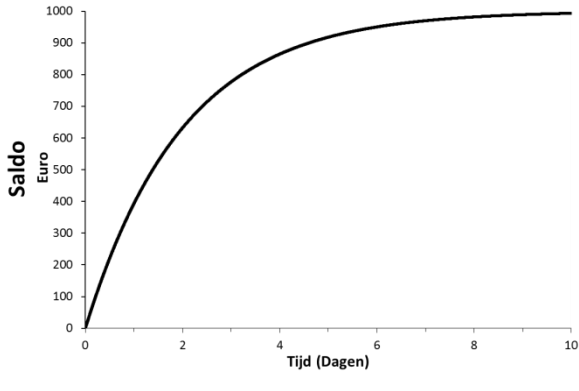
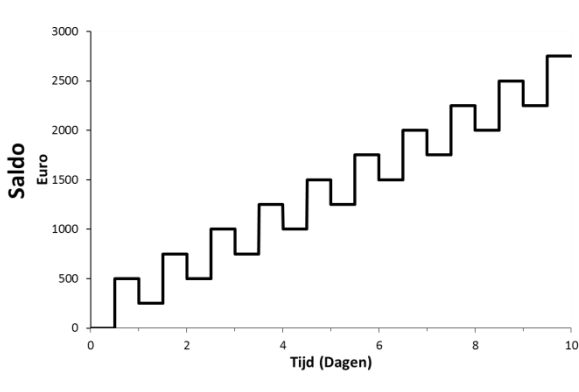
Question		Participants were asked to select the graph that best shows the mutation (net change) during the same time period as shown in the bank balance graph.	
A		Count (%)	Pattern
7 (8%)		A correlation heuristic response. A reverse image of the stock graph. Has little to no logical connection to the actual question except for the shape of the graph being similar.	
B		45 (54%)	The correct answer, it shows a constant net flow, positive when the stock is rising and negative when the stock is decreasing. The height of the net flow relates to the slope of the stock. The change in the stock corresponds to the area between the line plotting the net flow and the x-axis.
C		22 (27%)	A typical correlation heuristic response; a copy of the stock graph. Shows a failure to see the stock/flow relationship and erroneously assumes that the pattern of the net flow is equal to that of the stock. Has no connection to the material provided except for looking similar to the stock graph.

D		<p>4 (5%)</p>	<p>Although this graph is similar to the correct answers it switched the two ‘humps’ of the net flow. This results in showing changes in the behaviour of the net flow when no change in the behaviour of the stock occurs and vice versa. Furthermore the slopes of the two graphs do not correspond.</p>
E		<p>2 (2%)</p>	<p>Similar to the correct answer but fails to show that the decreasing stock value should relate to a negative value in the net flow, otherwise correct.</p>
F		<p>3 (4%)</p>	<p>Integration of the stock values. Results in an increasingly rising, linearly rising and decreasingly rising behaviour, corresponding to the linearly rising, constant, linearly decreasing behaviour of the stock value. Would be correct if the stock-flow relation would be inversed.</p>

## Appendix B – Behaviour from text and diagrams

### Detailed results of the bank balance task

<p>Question</p>		<p>Participants were asked to select the graph that best depicts the cash balance of an organization's bank account. The organization continuously receives deposits on this account and continuously expenses money. In total they receive 500 euro/day. They expense 50% of the balance per day. The initial balance is 0 euro. Participants received the question in text (and those in the manipulation condition also received the diagram on the left).</p>				
<p>A</p>		<table border="1"> <tr> <td data-bbox="896 853 1002 913">Count (%)</td> <td data-bbox="1002 853 1369 913">Pattern</td> </tr> <tr> <td data-bbox="896 913 1002 1240">2 (2%)</td> <td data-bbox="1002 913 1369 1240">This option shows the correct stock behaviour for a net flow of 500 euro/day. This ignores the outflow.</td> </tr> </table>	Count (%)	Pattern	2 (2%)	This option shows the correct stock behaviour for a net flow of 500 euro/day. This ignores the outflow.
Count (%)	Pattern					
2 (2%)	This option shows the correct stock behaviour for a net flow of 500 euro/day. This ignores the outflow.					
<p>B</p>		<table border="1"> <tr> <td data-bbox="896 1256 1002 1541">2 (2%)</td> <td data-bbox="1002 1256 1369 1541">Shows the stock in equilibrium. Has little to no connection to the information that was provided, other than misinterpreting the deposits per day as the initial cash balance.</td> </tr> </table>	2 (2%)	Shows the stock in equilibrium. Has little to no connection to the information that was provided, other than misinterpreting the deposits per day as the initial cash balance.		
2 (2%)	Shows the stock in equilibrium. Has little to no connection to the information that was provided, other than misinterpreting the deposits per day as the initial cash balance.					
<p>C</p>		<table border="1"> <tr> <td data-bbox="896 1659 1002 1944">4 (5%)</td> <td data-bbox="1002 1659 1369 1944">Another answer showing a misinterpretation of the inflow as the initial condition. It does show the correct outflow behaviour of 50% of the stock value per day, if the initial balance would have been 500 euro and there would have been no inflow.</td> </tr> </table>	4 (5%)	Another answer showing a misinterpretation of the inflow as the initial condition. It does show the correct outflow behaviour of 50% of the stock value per day, if the initial balance would have been 500 euro and there would have been no inflow.		
4 (5%)	Another answer showing a misinterpretation of the inflow as the initial condition. It does show the correct outflow behaviour of 50% of the stock value per day, if the initial balance would have been 500 euro and there would have been no inflow.					

D		<p>24 (27%)</p>	<p>Shows the stock behaviour for a net flow of 250 euro/day. Indicating that participants considered the expenses being 50% of the inflow and adding the resulting 250 euro/day to the stock.</p>
E		<p>36 (41%)</p>	<p>The correct answer. The constant inflow and balancing outflow results in goal seeking behaviour, in an upward direction, with the goal of a balance of 1000 euro. The goal is reached when inflow equals outflow. This happens when inflow = outflow = <math>.5 * \text{stock}</math>.</p>
F		<p>20 (23%)</p>	<p>A graph showing discrete behaviour and a failure to incorporate outflow behaviour depended on the stock value. 500 euro is added halfway each month and 50% of this value flows out at the end of the month.</p>

### Detailed results of the epidemic task

This task is based on the epidemic or SIR model as discussed in Sterman (2000).

<p>Question</p>		<p>Participants were asked to select the graph that best depicts the number of infected people. On day 1 one person is infected. The disease spreads through contact. A person only meets one other person each day. Sick and healthy people meet each other randomly. If a healthy person meets a sick person there is a 50% chance that the healthy person is infected. Participants received the question in text (and those in the manipulation condition also received the diagram on the left).</p>	
<p>A</p>		<p>Count (%)</p> <p>3 (3%)</p>	<p>Pattern</p> <p>Shows goal seeking behaviour generated by an information delay with a delay time of 4 days and a goal of 100 people. The curve is not based on any information that was provided and ignores the reinforcing feedback loop which is dominant in the first period.</p>
<p>B</p>		<p>0 (0%)</p>	<p>This graph shows the linear stock behaviour resulting from a constant flow of 9 persons/months. This value has no connection to the information provided.</p>

C		<p>26 (30%)</p>	<p>The correct response. At first a positive feedback loop is dominant causing people to be infected at an increasing rate. After some time many people are infected leading to a balancing loop to become dominant. This decreases the number of infections per day and results in goal seeking behaviour.</p>
D		<p>14 (16%)</p>	<p>This option displays exponential behaviour with a growth rate of 50% per day which stops when all are infected. This shows a misinterpretation of the flow equation and complete absence of the balancing loop.</p>
E		<p>1 (1%)</p>	<p>This graphs shows an increase, peak and decrease in the number of infected and looks similar to the flow behaviour (although faster). Ignores that there is no outflow. Little resemblance to the information provided.</p>
F		<p>44 (50%)</p>	<p>Shows the number of infected increasing at an increasing rate, with all people being infected at the final time. Ignores the balancing loop and uses an infection rate which is not related to the information provided.</p>