## 11: Ups and Downs! An Experimental Study into the Effect of Zigzag Shapes on Performance of the Department Store Task

*Hubert Korzilius, Tom Bongers and Stephan Raaijmakers*

### Introduction

Central in the system dynamics literature are the principles of accumulation and stock-flow reasoning. In system dynamics, every system is represented by a structure of stocks and flows. The inflow and outflow determine changes in the level of the stock. For example, in the problem of climate change, the stock of atmospheric $CO_2$ is increased by the inflow of anthropogenic $CO_2$ emissions and decreased by the outflow of $CO_2$ absorbed by oceans and biomass. To arrive at sustainable policy solutions it is necessary to have adequate systems thinking skills. Stock-flow tasks test if people can abstract the structure of a system based on its behavior and if people can reason in terms of stocks and flows. Research has shown that many individuals have trouble performing stock-flow (SF) tasks, such as the Department store task (Booth Sweeney & Sterman, 2000). As individuals have difficulties to understand the principle of accumulation, they often incorrectly use the correlation heuristic to solve SF tasks (Cronin, Gonzalez, & Sterman, 2009). When people use the correlation heuristic they incorrectly assume that the behavior of the behavior of the stock resembles the (net)flow. However, Korzilius, Raaijmakers, Rouwette, and Vennix (2014) show that individuals also experience other specific interpretation problems in making SF tasks, such as terminology used and the presentation of the graph. Hämäläinen, Luoma, and Saarinen (2013) state that this latter aspect, in their words the framing of the SF task, and much less a lack of understanding of accumulation, is responsible for the relatively bad performance on these tasks. In this research we, two colleagues and one former student of Vennix, present the findings of an experiment that contributes to this discussion by testing performance in the Department store task using a graphical display of zigzag shapes of inflows and outflows of people entering the store. This study therefore examined whether the shape of the curves used in the Department store task affects task performance and aims to contribute to insights into the problem of understanding accumulation in dynamic decision making. Findings are discussed in relation to existing research and avenues for further research are explored.

**Theoretical background**

SF tasks are embedded in the theory and methodology of system dynamics in which the behavior of complex systems is studied and simulated (Ford, 2010; Sterman, 2000). System dynamics was initiated by Forrester of the Massachusetts Institute of Technology to help managers understand industrial processes and systems. Today, system dynamics is more generally focused on understanding decision making when people are confronted with complex dynamic systems. The basic assumption of system dynamics is that the structure of the system drives its behavior. The structure is characterized by the following four hierarchical levels: 1) the closed boundary, 2) the feedback loop as the basic system component, 3) the levels (of stocks) and the rates (of flows), and 4) goals, observed conditions, discrepancy between goals and observed conditions and desired action (cf. Vennix, 2011, p. 111). Insight in the interplay of these characteristics is necessary to fully understand the behavior of the dynamic system (Cronin & Gonzalez, 2007; Ford, 2010; Forrester, 2009; Sterman, 2000), but in this paper we focus on the stocks and flows that guide accumulation as this a vital step for systems thinking.

An archetypical example of a stock is water in a bathtub. The water flows into the bathtub through the tap and flows out through the pipe into the drain. When during a time interval the amount of water flowing in exceeds the amount flowing out, the amount of water in the tub accumulates; the net flow > 0. This goes on until the bathtub overflows. It depends on the system boundaries how long this will take (e.g., size of the bathtub or rate of the flow) (Sterman, 2000). If the inflow of water is equal to the outflow in a time interval the stock is in balance; the net flow = 0. If the outflow is larger than the inflow the level of the stock decreases; the net flow < 0. Summarized, a "stock accumulates its inflows less its outflows, beginning with the initial value of the stock" (Sterman, 2000, p. 195).

The accumulation principle is a universal phenomenon that can be applied to all systems and is essential for comprehension and management of societal, corporate and individual decision making (Cronin et al., 2009). It is, for example, critical to understand the problem of climate change, where the stock of atmospheric $CO_2$ is increased by the inflow of anthropogenic $CO_2$ emissions and decreased by the outflow of $CO_2$ absorbed by oceans and biomass (Sterman, 2008). Also in people's daily life stocks and flows are important, for instance when managing one's bank account (stock) with deposits (inflows) and withdrawals (outflows) fluctuating over time (Cronin et al., 2009). In order to arrive at sustainable policy solutions (climate change) or make correct decisions (bank account) it is thus necessary to understand the complexity of dynamic systems. Therefore it is important that individuals have

adequate systems thinking skills among which understanding and being able to manage the accumulation principle (Hämäläinen et al., 2013).

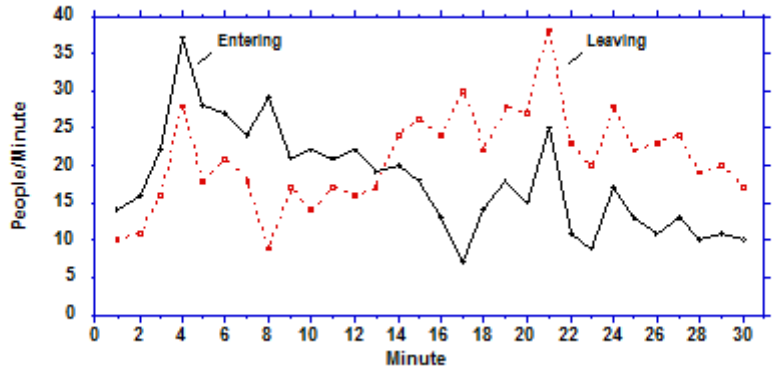
**Stock-flow (SF) tasks**

In order to investigate individuals´ system thinking skills, several SF tasks have been developed. Such tasks have in common that they present a dynamic problem after which participants need to answer a number of questions. In SF tasks, participants are presented a graph with inflows and outflows and, based on this information, have to determine the behavior of the stock while answering questions such as at which time the stock is at its maximum or minimum (e.g., Department store task in Figure 1)(Booth Sweeney & Sterman, 2000; Korzilius et al., 2014). Other tasks, in contrast, provide participants with information about a stock and ask them to estimate the net flows (Cash flow task; Veldhuis & Korzilius, in press). In a third category of tasks participants are not asked to estimate the stock or flows at a particular point in time, but are demanded to sketch the behavior of the stock or flows over time (such as the Bathtub task; Sterman, 2002). Often SF tasks are relatively simple containing one stock and one inflow and outflow. More complex tasks contain feedback loops and delays (e.g., the female professor task asks participants to bring two initial unequal stocks of female and male professors into balance; Bleijenbergh, Vennix, & Van Engen, 2011).

**Department store task**

One of the most often studied SF task is the Department store task (Sterman, 2002) (see Figure 1).

The graph below shows the number of people **entering** and **leaving** a department store over a 30 minute period.




Please answer the following questions.

Check the box if the answer cannot be determined from the information provided.

1. During which minute did the most people enter the store?

   Minute _____ ☐ Can't be determined

2. During which minute did the most people leave the store?

   Minute _____ ☐ Can't be determined

3. During which minute were the most people in the store?

   Minute _____ ☐ Can't be determined

4. During which minute were the fewest people in the store?

   Minute _____ ☐ Can't be determined

Figure 1. *Department store task (Sterman, 2002, p. 510)*

Figure 1 shows the relative simplicity of the Department store task. It focuses on accumulation and does not contain feedback mechanisms, delays, or non-linearity. The task presents a graph with two flows of people entering (inflow) and leaving (outflow) a department store during a 30-minute time interval, followed by four questions. Question 1 and 2 infer if participants can read the graph and correctly distinguish between the inflow and outflow (Cronin et al., 2009); the correct answers are minute 4 and 21, respectively. The other two questions assess whether individuals can deduce the behavior of the stock from the behavior of the flows (Cronin et al., 2009; Sterman, 2002). In order to solve these questions, the level of the stock at a specific time as well as the inflow and outflow rate have to be taken into account. Question 3 asks to indicate the highest level of the stock and Question 4 refers to the lowest level. For answering Question 3 it suffices to infer until what time the rate of people entering exceeds the rate of people leaving. The inflow exceeds the outflow (net flow $> 0$) until the graphs cross, so most people are in the department store at minute 13. After the intersection the outflow consistently exceeds the inflow (net flow $< 0$). In addition, the area between the curves after the intersection, is larger than the area before the intersection,

meaning that the total rate of leaving is greater than the total rate of entering. So the answer to Question 4, during which minute are the fewest people in the department store, is at the end, at minute 30 (Cronin et al., 2009; Sterman, 2002).

Research on the Department store task shows that many individuals, even highly educated, fail to correctly answer all four questions (Cronin et al., 2009; Sterman, 2010. This may implicate that: a) participants do not understand the acccumulation principle (Cronin et al., 2009), b) the problem representation of the accumulation principle is not optimal (Cronin & Gonzalez, 2007), and /or c) heuristic reasoning is triggered by the task (Hämäläinen et al., 2013). Inadequate problem representation may contribute to the complexity of the task, pushing as it were, to poor performance. On the other hand, particular features of the problem representation may also pull towards the use of specific heuristics.

Regarding problem representation, Cronin et al. (2009) showed that the finding of poor performance was stable in varying conditions and did not change performance: it appeared independent of cognitive burden (using fewer data points), graph display (presenting data in other formats, such as a table, text, or bar graph), task context (familiarity with context), receiving feedback (participants were given information which were answers were correct), motivation (informing participants that they could leave the experimental session once they had answered all questions correctly), and priming participants (of the presence and behavior of stock-flow structures). As a result of this (Cronin et al., 2009, p. 116) concluded that people fail to "appreciate the most basic principles of accumulation, leading to the use of inappropriate heuristics". However, according to Hämäläinen et al. (2013), the shape of graph may not only mask the accumulation principle but may also trigger people to use particular heuristics. In addition, they claimed that peaks and troughs selected in the graph are visually salient and therefore trigger the availability heuristic.

Kahneman (2011, p. 98) defines a heuristic as "a simple procedure that helps find adequate, though often imperfect, answers to difficult questions". A simple procedure refers to substituting a new, simpler question for the original, more difficult question. Related to SF tasks, Cronin et al. (2009, p. 124) state that the correlation heuristic, "a form of pattern matching in which people assume that the output of a system […] should ''look like" the input" is responsible for poor performance. Hämäläinen et al. (2013) state that the correlation heuristic is better covered by the well-known term availability heuristic, meaning that individuals make decisions based on information that is easiest to bring to mind, instead of exploring all pros and cons of plausible alternatives.

In a think aloud experiment, Korzilius et al. (2014) corroborated the prominent use of the correlation heuristic but also showed that participants have also other reasoning strategies while solving the Department store task. An example was the absence of explicit reasoning when performing the task. Another illustration was the incorrect assumption that, in order to determine the minute during which the most /fewest people were in the store(Questions 3 and 4), the initial value of the stock was needed. Participants also used a mix of the strategies mentioned above, which led to incorrect but also, in some cases, to correct answers. In addition, participants also expressed problems with reading the y-axis label containing a slash in the ratio people / minute, and with being unfamiliar with terminology used in the task.

**Department store task revised**

As argued above, the use of the correlation heuristic plays a role in why participants incorrectly solve the Department store task. Incorrect answers to Question 3 and Question 4 often fit with reasoning according to the correlation heuristic. Participants opt for the maximum in inflow or outflow (minute 4 and 21 in Figure 1), and particularly for the maximum difference (net flow) between inflow and outflow curves and vice versa (minute 8 and 17, Figure 1) as the correct answers to the question. According to Hämäläinen et al. (2013), these peaks and troughs are the most characteristic elements in the graph and therefore are more salient compared to other parts of the graph. As a result the presence of the peaks and troughs is more likely to induce erroneous reasoning. Hämäläinen et al. (2013, p. 626) contend "that in the department store task people's performance is affected by several cognitive heuristics triggered by a number of factors in the task that camouflage and divert people's attention from the true stock and flow structure".

As one of their experimental manipulations Hämäläinen et al. (2013) removed the peaks and troughs of the original Department store task, thereby removing the salient flow characteristics of the graph. In a series of four experiments using eleven different questionnaires they tested whether a revised graph with smoother curves resulted in better performance (see Figure 2). Although copying and pasting and the printing process may have been responsible, upon close observation the revised graph in Figure 2 seems to have two maxima in the entering line and it appears more difficult than in the original version to establish whether the area before the intersection is smaller than after the intersection (which is necessary for answering Question 4 of the original task).

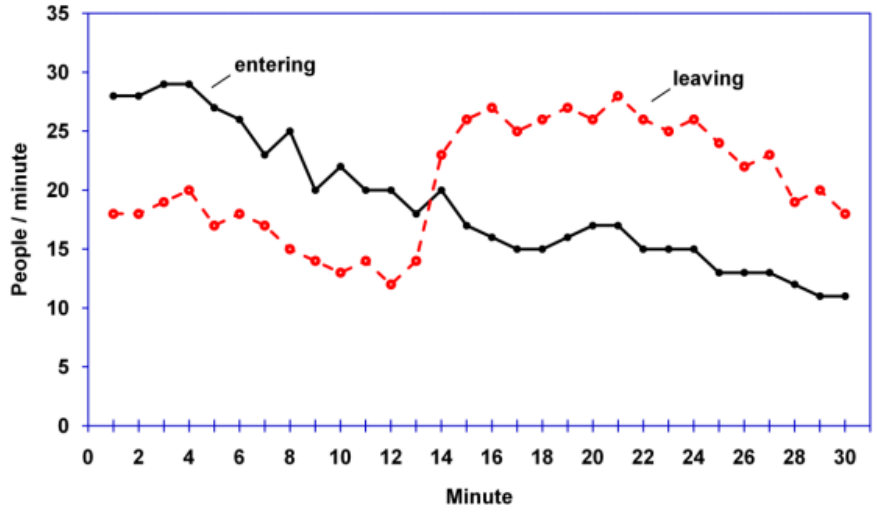



Figure 2. *Revised graph of Department store task using smoother curves (Hämäläinen et al., 2013, p. 629)*

Besides testing for *framing* the way the graph was presented as discussed above, Hämäläinen et al. (2013) also examined *priming* effects by varying the wording of the questions. They adapted the original wording by asking participants more directly about accumulation. "Q1. When did the number of people in the store increase and when did it decrease?" (p. 629). At the same time, they included additional elements: a "Cannot be determined" box and asking for a written explanation. In our view, these changes to the original task make it problematic to establish just the framing effect, thus isolating the effect of using smoother curves in comparison to the original curves.

In more detail: in their Questionnaire I (Hämäläinen et al., 2013, Table 1, p. 629) provided smooth curves. However, Hämäläinen et al. (2013) did not ask the original Question 1 and 2 (Cronin et al., 2009, Sterman, 2002). Instead they used the just quoted Q1 more straightforwardly focusing on accumulation. Next, they asked Question 3 and 4 of the original task but did not offer the "Cannot be determined" box. Together, differing curves, questions, and answering options make a fair comparison with performance on the original Department store task difficult.

Therefore, we think that Hämäläinen et al.'s (2013) claim "Our new results with somewhat revised experiments show that the poor performance in the department store task can be attributed to the framing of the problem rather than to people's poor understanding of the accumulation phenomenon" (p. 626) is too bold. This because it is not clear which adaptation, differently framing the graph or priming the questions and other elements, resulted in which

improvement of performance. To investigate the impact of graphical representation on performance, one has to rule out all other factors that might influence this relation.

**Department store task zigzagged**

Notwithstanding our comments on the study of Hämäläinen et al. (2013) we endorse their plea for more insight in and explanations of SF performance, such as the influence of graphical representation of information on stock-flow performance. Ultimately aiming to contribute to more knowledge of systems thinking as a vital part of system dynamics research. We tested whether heuristic reasoning is triggered by characteristics of the graph keeping all other elements of the problem formulation similar. Following Hämäläinen et al. (2013) we wanted to distract attention away from the few chacteristic points of the original Department store task (Figure 1). However, instead of using smoother curves (Figure 2), we designed the graph in such a way that it had even more peaks and troughs ('Ups and Downs'; see Figure 3) than the original version. We substantiated this adjustment by the argument that the visibility of the flow characteristics can be reduced, not only by scaling down the peaks in the graph (especially t4, t8, t17, and t21), but also by enlarging the contrasts in the rest of the graph. Therefore, we assumed that presenting more instances of net flow differences (inflow-outflow or vice versa) than in the original task would reduce the extent to which participants use the correlation heuristic. If this would be evidenced, the implemented adjustments apparently contribute to the internal validity of the task. Consequently, we formulated the following hypothesis:

Hypothesis 1. An articulated zigzagged version of the Department store task will result in less correlation heuristic reasoning than the original version.

In addition to this, although more difficult to substantiate, we assumed that in real life peaked curves are more common than smooth curves for illustrating dynamic behavior, for example curves used for stock markets and weather forecasts. A Google search using the search term "line graph with two lines" corroborated this as it resulted in numerous irregular, rather than smooth curves. Consequently, zigzagged curves may be more familiar and thus may promote external validity of the graph. These considerations resulted in the following hypothesis:

Hypothesis 2. An articulated zigzagged version of the Department store task will perform better than the ones who get the original version.

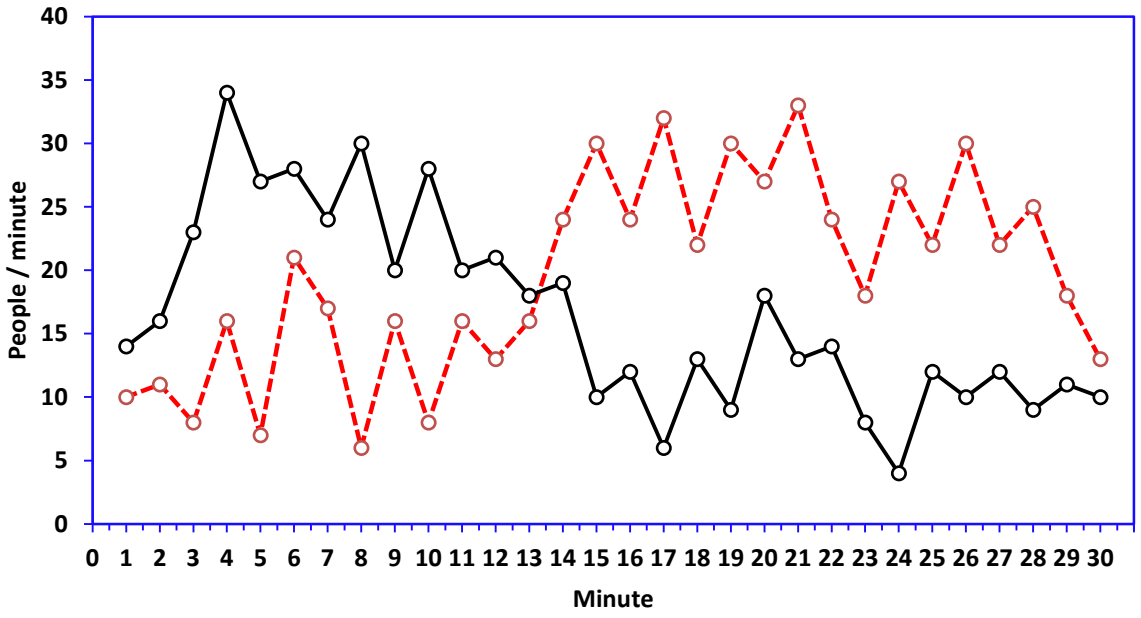



Figure 3. *Articulated zigzagged shape of graph Department store task used in this study*

## Method

*Experimental design and procedure*

In line with previous research (Cronin et al., 2009; Korzilius et al., 2014; Sterman, 2010) we tested the effect of graphical representation on performance and correlation heuristic reasoning, using a one-factorial randomized between-subjects experimental design. Participants were randomly assigned to a version of the Department store task. Participants in the Experimental group received a zigzagged graph (Figure 3), while those in the Control group got the graph of the original task (Figure 1).

Participants in both groups had to answer the same four questions as in the original task:

1. During which minute did the most people enter the store?
2. During which minute did the most people leave the store?
3. During which minute were the most people in the store?
4. During which minute were the fewest people in the store?

Likewise, the same answering options as in the original Department store task were used: either fill in the minute or check the box "Can't be determined" (see Figure 1). To rule out the possible influence of proficiency in English, we translated both versions in Dutch.

The shape of the inflows and outflows in the zigzagged graph contained more peaks and troughs, and thus more instances of larger net flow differences than the original graph of the Department store task. Because the graph contained more peaks and troughs, the absolute

values of the stock and flow of the zigzagged graph at the various minutes differed from the original graph. However, important for comparison purposes, the minutes at which the stock and flows reached important values were kept similar to the original. This meant that the maximum entering/leaving the store and maximum/fewest in stock were the same as in the original task, minute 4, 21, 13, and 30, respectively. Also, indicative for the correlation heuristic, max net inflow/outflow was the same, minute 8 and 17, respectively. Further, in line with Cronin et al. (2009, p. 118, note 3) the design of the zigzagged graph was such that the area of the region before the intersection, where the inflow is larger than the outflow, is clearly smaller than the area after the intersection where the outflow is bigger than the inflow. Finally, the layout was similar to the original task with one exception. In order to facilitate reading, we provided all minutes on the x-axis instead of even minutes only. This was done in the experimental and control group in the same way.

*Procedure and participants*

The experiment was conducted at the office of a Dutch based international staffing agency in the catering industry. This particular population was chosen for their expected high homogenitey and for the possibility of finding a large group of participants as one of the authors was in the management team of the company. Data collection took place shortly before employees had to start or had finished their work. Participation was voluntary and no reward was offered. Participants were not allowed to use a computer or calculator and had a maximum of 10 minutes to make the task.

Participants were 76 employees, 60.0% male, on average 22.6 years old (range 18-32), mostly students working part-time for a Dutch based staffing agency in the catering industry. Table 1 shows the characteristics of the participants. The Experimental group consisted of 41 participants, the Control group of 35. The majority of participants stated not to have much knowledge of System Dynamics. They were in general higher educated in the fields of Management, Behavior and society or Law.

Table 1. *Characteristics of participants in Experimental Group (EG) and Control group (CG)*

| | EG | | CG | | Total | |
|---|---|---|---|---|---|---|
| | Zigzag shape | | Original shape | | | |
| | 41 | (53.9) | 35 | (53.9) | 76 | (100.0) |
| Age | 23.0 | (2.85) | 22.2 | (2.02) | 22.6 | (2.52) |
| Gender | | | | | | |
| Male | 23 | (56.1) | 22 | (64.7) | 45 | (60.0) |
| Female | 18 | (43.9) | 12 | (35.3) | 30 | (40.0) |
| Knowledge of system dynamics | | | | | | |
| Very little | 19 | (47.5) | 16 | (48.5) | 35 | (47.9) |
| Little | 10 | (25.0) | 12 | (36.4) | 22 | (30.1) |
| Not little, not much | 11 | (27.5) | 5 | (15.2) | 16 | (21.9) |
| Level of completed education | | | | | | |
| Primary | 0 | (0.0) | 2 | (5.9) | 2 | (2.7) |
| Secondary | 6 | (14.6) | 11 | (32.4) | 17 | (22.7) |
| Intermediate Vocational | 2 | (4.9) | 1 | (2.9) | 3 | (4.0) |
| University of Applied Sciences | 3 | (7.3) | 0 | (0.0) | 3 | (4.0) |
| University Propaedeutic | 13 | (31.7) | 11 | (32.4) | 24 | (32.0) |
| University Bachelor | 13 | (31.7) | 5 | (14.7) | 18 | (24.0) |
| University Master | 4 | (9.8) | 4 | (11.8) | 8 | (10.7) |
| Field of current education | | | | | | |
| Management | 14 | (35.0) | 5 | (14.7) | 19 | (25.7) |
| Behavior and society | 8 | (20.0) | 6 | (17.6) | 14 | (18.9) |
| Law | 4 | (10.0) | 6 | (17.6) | 10 | (13.5) |
| Other | 12 | (30.0) | 12 | (35.3) | 24 | (32.4) |
| No | 2 | (5.0) | 5 | (14.7) | 7 | (9.5) |

*Note*. Cell entries indicate *n*s and % between brackets; except for Age reporting M (SD).

An independent *t*-test (for Age) and Chi-square analyses (for the other) revealed no differences in background characteristics reported in Table 1 between the Experimental and Control group.

A power analysis (G*Power Version 3.1.92) showed that with the number of participants per group, we anticipated to find medium to large differences between the two groups (effect size = 0.58) in 80% of the cases (statistical power = .80) conducting one-tailed *t*-tests at an alpha level of .05 (Cohen, 1992).

*Measures and statistical analyses*

The variable *group* represented the manipulation of the experiment containing the experimental group, having the zigzagged version of the Department store task, and the control group, having the original version.

*Performance per question* was measured in terms of either correctly answering or not correctly answering Question 1 to 4, with correct answers being minutes 4 (Q1), 21 (Q2), 13 (Q3), and 30 (Q4), respectively. Additionally, to establish performance for the questions in which accumulation was involved, *performance total* was computed by adding the number of correct answers for Question 3 and 4; theoretical range 0-2.

The measurement of the correlation heuristic reasoning was also based on Question 3 and 4. *Correlation heuristic per question* was measured in terms of either answering minute 8 (max net inflow) to Question 3 and minute 17 (max net outflow) to Question 4. Also, *correlation heuristic total* was computed by adding the number of correlation heuristic answers; theoretical range 0-2.

SPSS Version 22 was used to conduct the statistical analyses. To compare the two groups on the variables Performance and Correlation heuristic per question, Chi-square tests were used. In line with the direction of the hypotheses, one-sided independent *t*-tests were conducted to test the effect of the task on the variables Performance total and Correlation heuristic total. It appeared that all outcomes of the parametric t-tests were corroborated by the non-parametric alternative Mann-Whitney tests, therefore, we only present parametric outcomes.

Beyond the effect of group, we analyzed effects of control variables, by Analyses of Covariance (ANCOVA; control variables age and level of completed education) and by factorial two-way analyses of variance (ANOVA; other control variables). We limited these analyses to the dependent variables correlation heuristic total and performance total.

The alpha level for all tests was set at .05.

## Results

*Descriptives*

Table 2 shows the performance on the Department store task of the participants in the experimental and the control group. It reveals similar patterns of task-flow performance as

reported in previous research (e.g., Cronin et al., 2009; Korzilius et al., 2014; Sterman, 2002, Pala & Vennix, 2005). Participants generally did not have problems answering Question 1 and Question 2. The percentages in the underlined cells in the columns of Question 3 and 4 indicates that quite some participants used correlation heuristic reasoning, and that, especially for Question 3, the relative frequency was higher in the experimental group than in the control group. The limited number of correct answers of Question 3 and Question 4, demonstrate that participants in both groups had difficulties with the concept of accumulation. For answering the last two questions, relatively many participants opted for "Can't be determined".

*Testing hypotheses*

*Hypothesis 1*. There appeared no difference in the variable Correlation heuristic between the experimental and control group in Question 3 ($\chi^2(1, n = 76) = 0.25$ $p = .62$), Question 4 ($\chi^2(1, n = 76) = 0.01$, $p = .95$), nor for Correlation heuristic total ($M_{\text{experimental group}} = 0.34$, $SD_{\text{experimental group}} = 0.66$; $M_{\text{control group}} = 0.57$, $SD_{\text{control group}} = 0.78$; $t(74) = 1.40$, $p = .083$, one-sided). Although the descriptive statistics may have pointed to a possible difference, Hypothesis 1 was rejected.

*Hypothesis 2*. There were no differences between the participants in the experimental and the control group for the four separate questions of the Department store task (Question 1: $\chi^2$ $(1, n = 76) = 2.67$, $p = .10$; Question 2: $\chi^2$ $(1, n = 76) = 0.95$, $p = .33$; Question 3: $\chi^2(1, n = 76)$ $= 0.25$ $p = .62$; Question 4: $\chi^2(1, n = 76) = 0.01$, $p = .95$). Performance total was also not statistically different ($M_{\text{experimental group}} = 0.61$, $SD_{\text{experimental group}} = 0.83$; $M_{\text{control group}} = 0.66$, $SD_{\text{control group}} = 0.87$; $t(74) = 0.24$, $p = .20$, one-sided). Accordingly, Hypothesis 2 was rejected. This means that adaptation of the original curve of the Department store task into a zigzagged curve did not have any effect on the use of the correlation heuristic nor on the performance of the task.

Although there was no evidence for the hypotheses, we additionally performed analyses of control variables (age, gender, knowledge of System Dynamics, level of completed eduction, and field of education) to explore whether they might have had an effect. This was not the case except that Level of completed education was negatively related to correlation heuristic total ($r_s = -.31$, $p < .01$).

Table 2. *Results Department store task for Experimental group (EG) and Control group (CG)*

| Answers | Question 1 | | | | Question 2 | | | | Question 3 | | | | Question 4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Most entering | | | | Most leaving | | | | Most in store | | | | Fewest in store | | | |
| | EG | | CG | | EG | | CG | | EG | | CG | | EG | | CG | |
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| Max entering *t* = 4 | **38** | **92.7** | **35** | **100** | | | | | 1 | 2.4 | | | | | | |
| Max leaving *t* = 21 | | | | | **35** | **85.4** | **33** | **94.3** | | | 2 | 5.7 | | | 1 | 2.9 |
| Max in stock *t* = 13 | | | | | | | | | **13** | **31.7** | **12** | **34.3** | 4 | 9.8 | 4 | 11.4 |
| Fewest in stock *t* = 30 | | | | | | | | | | | | | **12** | **29.3** | **9** | **25.7** |
| Max net inflow *t* = 8 | 2 | 4.9 | | | 2 | 4.9 | | | 8 | 19.5 | 12 | 34.3 | | | | |
| Max net outflow *t* = 17 | | | | | 1 | 2.4 | | | 2 | 4.9 | 2 | 5.7 | 6 | 14.6 | 8 | 22.9 |
| Initial in store *t* = 1 | | | | | | | | | | | | | | | 2 | 5.7 |
| Can't be determined | 1 | 2.4 | | | 1 | 2.4 | | | 13 | 31.7 | 5 | 14.3 | 11 | 26.8 | 6 | 17.1 |
| Other | | | | | 2 | 4.9 | 2 | 5.7 | 4 | 9.8 | 2 | 5.7 | 8 | 19.5 | 5 | 14.3 |
| No answer | | | | | | | | | | | | | | | | |

*Note*. EG (*n* = 41) had the zigzagged version (see Figure 3), CG (*n* = 35) the original version (see Figure 1).

The rows are the answers with the time point indicated in column 1 (answers to all questions were considered correct if they were within 1 minute of the correct response). Conform Cronin et al. (2009, p. 119), bold numbers indicate correct responses; underlined numbers show the incorrect, correlation heuristic, answers for Question 3 and 4 that give the maximum net inflow/net outflow instead of maximum/fewest in the stock.

**Conclusion and discussion**

We aimed to contribute to the understanding of accumulation by conducting an experiment in which we tested the effect of graphical representation on performance in the Department store task. We examined whether a graphical respresentation presenting inflows and outflows in an articulated zigzagged shape would do better than the original graph (Cronin et al., 2009; Sterman, 2002). We expected that a zigzagged graph would draw attention away from the few typical characteristics of the original graph and as a result would reduce correlation heuristic reasoning and increase performance. Although there appeared fewer instances of correlation heuristic reasoning in the experimental group having the zigzagged graph than in the control group having the original graph, especially while answering Question 3, the differences were not statistically significant. Hypothesis 1, stating that an articulated zigzagged version of the Department store task leads to less correlation heuristic reasoning than the original version, was therefore rejected. Hypothesis 2 was also rejected: Contrary to our expectations, participants assigned to the articulated version of the Department store task did not perform better than participants confronted with the original version. Based on the outcomes of this study we conclude that a graphical articulation of in- and outflows does not affect heuristic reasoning and performance.

Cronin et al. (2009) launched the correlation heuristic in their effort to understand the main pattern of answers given in the Department store task. Strictly speaking however, correlation reasoning, comprehended by them as the substitution of flow features for stock characteristics, is not an explanation but rather a description of what actually takes place. Although this descriptive knowledge has been corroborated in many studies, it does not *explain* why individuals seem to use correlation reasoning (see MacDonald Ross, 2001). Hämäläinen et al. (2013) did search for an explanation of correlation reasoning in the *availability* of particular graph characteristics. They smoothed the peaks and troughs of the original Department store task to reduce availability. Unfortunately, the claims about their research findings were undermined by shortcomings in their experimental design. In the current study we followed the approach of Hämäläinen et al. and complemented it by using a graph with an articulated zigzag pattern. We assumed that presenting more instances of net flow differences would also reduce the availability of the original flow characteristics and therefore would lead to less correlation reasoning and better performance. However, our expectations were not evidenced. Future research on description and explanation of heuristics is therefore necessary to eventually grasp why individuals have poor performance on SF

tasks. In general, in line with the initiative of the Open Science Collaboration (2015), we encourage more replication of experiments on accumulation. As they state: “Scientific claims should not gain credence because of the status or authority of their originator but by the replicability of their supporting evidence” (p. 943). Although research inevitably has its ups and downs, the spirit that emerges from this quotation is exactly in line with the attitude of the Methodology group at Radboud University in Nijmegen, initiated by Jac Vennix.

**References**


Bleijenbergh, I.; Vennix, J.; Jacobs, E.; van Engen, M. (2011). Reducing the gender gap: Biases in understanding delays in personnel policies. *In Proceedings of the 29th International Conference of the System Dynamics Society, Washington D.C., USA.* System Dynamics Society: Albany, NY.

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

Cohen, J. A. (1992). Power Primer. *Psychological Bullitin, 122*, 155-159.

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

Cronin, M., Gonzalez, C., & Sterman, J. D. (2009). Why don't well-educated adults understand accumulation? A challenge to researcher, educators and citizens. *Organizational Behavior and Human Decisision Processes*, *108*, 116-130.

Ford, A. (2010). *Modeling the Environment* (2nd ed.). Washington DC: Island Press.

Forrester, J. (2009). *Some basic concepts in System Dynamics.* Sloan School of Management. Waltham, MA: Massachusetts Institute of Technology.

Hämäläinen, R. P., Luoma, J., & Saarinen, E. (2013). On the importance of behavioral operational research: The case of understanding and communicating about dynamic systems, *European Journal of Operational Research*, *228*, 623–634.

Kahneman, D. (2011). *Thinking, fast and slow.* London (etc.): Allen Lane.

Korzilius, H., Jong, E., de, & Raaijmakers, S. (2015). A different outlook on stock-flow tasks. Using eye tracking methodology to explore eye movements of problem solvers. *Presentation at 27th European Conference on Operational Research*, Glasgow, 12-15 July.

Korzilius, H., Raaijmakers, S., Rouwette, E., & Vennix, J. (2014). Thinking aloud while solving a stock-flow task: Surfacing the correlation heuristic and other reasoning patterns. *Systems Research and Behavioral Science*, *31,* 268-279.

MacDonald Ross, G. (2001). *Leibniz*. Rotterdam: Lemniscaat.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). DOI: 10.1126/science.aac4716

Pala, O., & Vennix, J. (2005). Effect of system dynamics education on systems thinking inventory task performance. *System Dynamics Review*, *21*, 147-172.

Sterman, J. D. (2000). *Business Dynamics. Systems thinking and modeling for a complex world.* Boston, MA: McGraw-Hill.

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

Sterman, J. D. (2008). Risk communication on climate: Mental models and mass balance. *Science*, *332*(5901), 532-533.

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

Veldhuis, G., & Korzilius, H. (in press). Seeing with the mind. The relationship between spatial ability and inferring dynamic behavior from graphs. *Systems Research and Behavioral Science.*

Vennix, J. A. M. (2011). *Theorie en praktijk van empirisch onderzoek* [Theory and practice of empirical research] (5th ed.). Harlow: Pearson Education.