Systems Research and Behavioral Science
Syst. Res. **31**, 268–279 (2014)
Published online 17 June 2013 in Wiley Online Library
(wileyonlinelibrary.com) **DOI**: 10.1002/sres.2196

■ *Research Paper*

Thinking Aloud While Solving a Stock-Flow Task: Surfacing the Correlation Heuristic and Other Reasoning Patterns

Hubert Korzilius*, Stephan Raaijmakers, Etiënne Rouwette and Jac Vennix

Institute for Management Research, Radboud University Nijmegen, the Netherlands

In the literature, it is assumed that individuals, while performing stock-flow tasks, often use a correlation heuristic, a form of pattern matching in which they think that the behavior of the stock resembles the (net) flow. To investigate this assumption and to increase our insight in the actual reasoning patterns when performing stock-flow tasks, we conducted an experiment by using the department store task as baseline. In the treatment condition, participants performed the stock-flow task while thinking aloud; in the control condition, they only had to write down their answers. The correlation heuristic was corroborated: participants actually did verbalize their thoughts in terms of the biggest difference between inflow and outflow at a particular point, thus expressing the correlation heuristic in words. However, other reasoning strategies that led to incorrect claims were also found. Further research is desirable to elaborate insight in the precursors of heuristic reasoning.

Keywords stock-flow performance; think aloud method; reasoning patterns; correlation heuristic

INTRODUCTION AND THEORETICAL BACKGROUND

In the tradition of research on stock-flow task performance (Booth Sweeney and Sterman, 2000; Cronin *et al*., 2009; Sterman, 2010), this study aims to shed light on the way accumulation processes are understood. The influence of task context is tested along with the correlation heuristic as a problem-solving strategy. Insight in the use of the correlation heuristic is gained by employing the think aloud method, a method that enables us to reveal reasoning patterns in problem-solving situations.

Research on dynamic decision making in which individuals perform stock-flow tasks shows that the principle of accumulation is not well understood and that individuals often seem to use the wrong heuristics while completing such assignments (Booth Sweeney and Sterman, 2000; Pala and Vennix, 2005; Cronin *et al*., 2009;

*Correspondence to: Hubert Korzilius, Institute for Management Research, Radboud University Nijmegen, the Netherlands
E-mail: h.korzilius@fm.ru.nl

Received 29 April 2013
Accepted 30 April 2013

Sterman, 2010). A variety of tasks have been used in these experiments, ranging from very simple structures with one stock with only one inflow to more complex ones with two or more stocks with corresponding inflows and outflows (e.g. Booth Sweeney and Sterman, 2000; Kainz and Ossimitz, 2002; Ossimitz, 2002; Sterman, 2002). All of the experiments reveal that most individuals, even highly educated ones, have difficulties with the most simple stock-flow tasks (e.g. Booth Sweeney and Sterman, 2000). Moreover, training of individuals in stock-flow thinking appears to be only partly successful (Kainz and Ossimitz, 2002; Pala and Vennix, 2005).

Research also demonstrates that task context in terms of salience and familiarity does not have an effect on stock-flow task performance, nor does the ability to interpret graphs (differences between deriving information from line graphs, tabular format, bar graphs and text), nor cognitive capacity (the number of data points in the graph; Booth Sweeney and Sterman, 2000; Pala and Vennix, 2005; Cronin *et al*., 2009; Brunstein *et al*., 2010; Sterman, 2010). Similarly, reflection on cognitive conflict as a source to advance conceptual change (generated by utilizing a running total calculation) does not improve individuals' performance (Phuah, 2010).

In general, when facing complex problems or incomplete information, individuals often draw on heuristic reasoning, replacing the original question by a related question that is easier to answer (Tversky and Kahneman, 1974; Kahneman *et al*., 1982). In the context of stock-flow problems, Cronin *et al*. (2009) coin the term correlation heuristic, as a form of pattern matching in which individuals assume that the behavior of the stock resembles the (net) flow (cf. Booth Sweeney and Sterman, 2000; Sterman, 2010).

Thus far, research on stock-flow performance has focused on the outcomes of reasoning processes and inferred that individuals use correlational reasoning while estimating stock-flow behavior, assuming that the flow(s) immediately and directly affect the stock. The actual reasoning process of participants remained hidden from the researchers. As a result, little is known about how the participants experience their performance and how they describe their thinking patterns or strategies. It is also not clear whether they in some sense touch upon the principle of accumulation. Another matter is whether participants themselves recognize and solve the task corresponding to the correlation heuristic. We may say that the correlation heuristic has the status of a hypothetical idea, a presumption that still has to be tested in research. Therefore, and in line with the recommendations of Pala and Vennix (2005) and Sterman (2010), the current study uses the think aloud method to increase insight in this matter. 'Thinking aloud during problem-solving means that the subject keeps on talking, speaks out loud whatever thoughts come to mind, while performing the task at hand' (Van Someren *et al*., 1994, p. 25). Thinking aloud thus helps us to understand how an answer is reached, explicating the reasoning patterns while individuals try to solve a problem or task.

This paper investigates the reasoning patterns used by participants in an experiment on the department store task, and a comparable newly developed bank task to check for task context effects. With this, we check whether the findings on task context effects of Cronin *et al*. (2009) could be replicated. To test if thinking aloud does not interfere with the outcome (principle of nonreactivity), we compared the results of participants performing the task in two different conditions: thinking aloud and writing (see section on Design and Procedure).

METHOD

Participants

Participants were 115 second year students of Business Administration of Radboud University Nijmegen in the Netherlands, following a Management game course. Students were told that as part of the course, a short exercise on decision making had to be performed. Participation was compulsory and a reward (a cake) was given to the best scoring participants. There were 53 (46%) female and 62 (54%) male participants. Their mean age was 20.6 years (range 19–27; standard deviation = 1.41). Three quarters of the

Syst. Res. **31**, 268–279 (2014)
DOI: 10.1002/sres

students ($n = 87$; 76%) had, as prior high school profile, Economics and Society,[1] the majority ($n = 111$; 97%) had the Dutch nationality, and four students (3%) had played the Beer game (Sterman, 1989) before.

THINK ALOUD METHOD

We applied the think aloud method as a valid approach to reveal the reasoning patterns of the participants (Van Someren *et al*., 1994). Unlike approaches where participants are asked to express and explain their thoughts after task performance, the think aloud method does not suffer from the risk that not all information is retrieved, or even worse, that false memories are reported (Ericsson and Simon, 1993; Van Someren *et al*., 1994). A meta-analysis (Fox *et al*., 2011) shows that the think aloud method is nonreactive, essentially meaning that data collection through verbalization of thoughts does not interfere with the outcomes of the task (cf. Ericsson and Simon, 1993). To give a concurrent account of one's thoughts seems harder than it is. 'For most people speaking out loud their thoughts becomes a routine in a few minutes' (Van Someren *et al*., 1994, p. 26). Thus, thinking aloud can be considered as an efficient and valid way of collecting information from the participants performing the task.

DESIGN AND PROCEDURE

In the experiment, we used a 2×2 factorial design in which participants had to complete a stock-flow task and were randomly assigned to two factors. The first factor differentiated between participants that either orally verbalized their argumentation while performing the task or did not explicitly verbalize their reasoning strategy. Factor 1 was labelled as *mode*, with conditions, think aloud ($n = 50$; 43%) or written ($n = 65$; 57%). The second factor referred to the task that participants had to complete, either the department store task or the, newly developed, bank task (see section on Materials and Measures). Factor 2 was named *task context*, with the following conditions: department store ($n = 61$; 53%) and bank task ($n = 54$; 47%). A power analysis revealed that with 115 participants, we expect to find large to medium differences between the groups (effect size $f = 0.27$, which is almost medium), with the statistical tests used at an alpha level of 0.05 for 80% of the cases (statistical power = 0.80; Cohen, 1992). Table 1 shows the distribution of the participants over the experimental conditions.

The procedure used for the stock-flow task was the following. Participants arriving at the experimentation rooms were welcomed by a research assistant who informed them that they were about to perform a task about a department store/bank. In the think aloud condition, each participant, sitting alone in a room, was addressed by the research assistant as follows (bank task):

> You are requested to complete a bank task that involves deposits and withdrawals at the bank. These are depicted in a graph and we ask you to answer four questions.[2] We also ask you to think out loud while you are answering the questions. In order not to influence you I will be quiet most of the time, only now and then I will encourage you to think aloud. We will make a video tape for research purposes only and make sure this remains anonymous. You are requested not to inform others of your solution. In total you have 10 minutes. Good luck!

In the think aloud condition, we were primarily interested in the reasoning strategies applied to solve the third and fourth problem questions. Nevertheless, we asked the respondents to think out loud right from the start. Answering the first and second questions served as a warm-up task, in which participants were comparatively easily familiarized with thinking aloud (cf. Ericsson and Simon, 1993).

In the written condition, two or three participants simultaneously performed the task in one room at detached seats under surveillance of a research

[1] A high school profile that is characterized by the compulsory subjects Economics and History, and with Management & Organization, Geography and Social sciences as electives. The other high school profiles were as follows: Culture and Society (3%), Science and Health (11%), Science and Technique (3%) and other profile (6%).

[2] These are questions 1 to 4 in Figure 1.

Syst. Res. **31**, 268–279 (2014)
DOI: 10.1002/sres

Table 1 Breakdown of participants over experimental conditions

		Mode		
		Think aloud	Written	Total
Task context	Department store	27	34	61
	Bank	23	31	54
	Total	50	65	115

assistant. They were told that they had to complete the task individually and were not allowed to talk to each other. The other instructions were the same as in the think aloud condition. The written condition was not video taped.

The department store and bank task were pre-tested by two university lecturers and two Master students. The data were collected in 2010.

MATERIALS AND MEASURES

In line with existing research (e.g. Pala and Vennix, 2005; Cronin *et al.*, 2009), an English version of the department store task was used. In addition, to examine whether there were task context effects, we developed an alternative stock-flow task we refer to as the bank task (Figure 1). For the Business administration students in the sample, we deemed financial inflows and outflows of a bank more salient and familiar than the flows of people into and out of a store (Cronin *et al.*, p. 121). We closely followed the format of the department store task regarding the layout, number of words and the wording of the four questions.

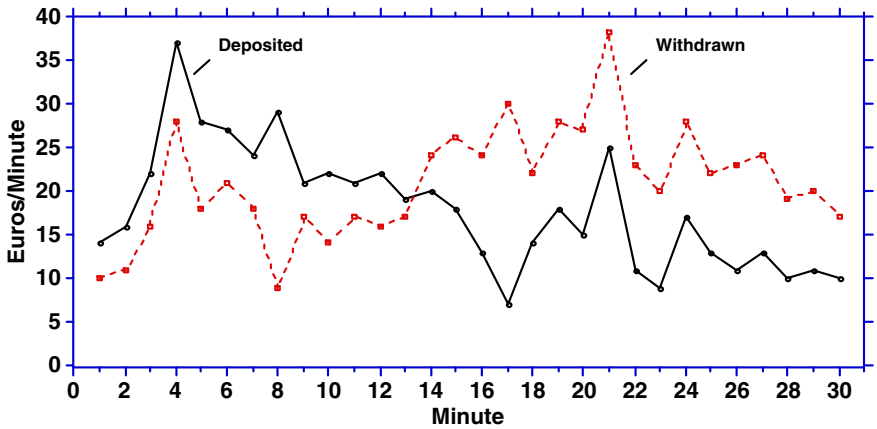
We established if the answers of the four questions were correct or incorrect. In line with previous research (e.g. Cronin *et al.*, 2009), we considered answers to all questions correct if they were within 1 min of the correct response. In addition, we assessed the total number of correct answers (theoretical range 0–4).

We transcribed the video tapes of 44 participants (because of technical shortcomings, six of the fifty recordings could not be used) and focused on the reasoning used for answering questions 3 and 4 (Q3 and Q4). In line with coding procedures for the analysis of qualitative data (Strauss and Corbin, 2007; Boeije, 2010), the content of each transcription was searched for keywords that revealed the internal structure of the participants' reasoning or the absence of such a structure. Some examples are as follows: *most/fewest Euros/people*, *difference (is biggest/smallest)*, *higher/lower than*, *more entered/entering (than) versus more left/leaving (than)*, *more deposited/depositing (than)*, *versus more withdrawn/withdrawing (than)*, *increase/increasing versus decrease/decreasing*. To prevent the risk of registering only terms anticipated, we also explicitly searched for keywords falling outside a strict system dynamics frame of reference (e.g. *take an intersection point*; *no exact numbers*). These keywords and the phrases they are part of refer to stocks or flows more or less explicitly and more or less accurately. Next, these keywords and key phrases were characterized in terms of stock-flow reasoning lines (A through M). Subsequently, these characterizations were categorized into more abstract reasoning types (1 through 7; see Table 4 for reasoning lines and reasoning types). In formulating the reasoning lines, an iterative procedure was followed, moving back and forth between the transcriptions, resulting in a preliminary categorization and allocation of the participants.

This initial search for the reasoning strategies of the participants was performed by the second author. Next, the third author independently coded all 44 protocols, so that we were in a position to assess interrater reliability. Regarding the allocation of the participants, there was 79% agreement about the reasoning lines used for Q3 and Q4, Cohen's kappa was 0.69. For reasoning types, these figures were 83% and 0.73, respectively. Any differences were subsequently settled by consent resulting in the final classification and allocation reported in the paragraph sections that follow.

Syst. Res. **31**, 268–279 (2014)
DOI: 10.1002/sres

The graph below shows the number of Euros being **deposited** in and **withdrawn** from a bank 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 were the most Euros deposited in the bank?

   Minute _____ ☐ Can't be determined

2. During which minute were the most Euros withdrawn from the bank?

   Minute _____ ☐ Can't be determined

3. During which minute were the most Euros in the bank?

   Minute _____ ☐ Can't be determined

4. During which minute were the fewest Euros in the bank?

   Minute _____ ☐ Can't be determined

Figure 1 The bank task

RESULTS

We checked whether there were differences in the performance of the department store and the bank task (factor 2 task context). This appeared not to be the case, not for answers to the four questions following the task (Q1: $\chi^2(1, n = 115) = 0.23$, $p = .63$; Q2: $\chi^2(1, n = 115) = 0.04$, $p = .84$; Q3: $\chi^2(1, n = 115) = 0.14$, $p = .71$; and Q4: $\chi^2(1, n = 115) = 0.07$, $p = .79$), nor for the total number of correct answers ($\text{mean}_{\text{dept.store}} = 2.23$, standard $\text{deviation}_{\text{dept.store}} = 0.78$; $\text{mean}_{\text{bank}} = 2.22$, standard $\text{deviation}_{\text{bank}} = 0.82$; $t(113) = 0.05$, $p = .96$). On average, participants needed 5 min to complete Q3 and Q4 (standard deviation = 2 min).

We also controlled whether the think aloud and written condition (factor 1 mode) yielded different results. This appeared not to be the case (Q1: $\chi^2(1, n = 115) = 0.13$, $p = .72$; Q2: $\chi^2(1, n = 115) = 0.05$, $p = .82$; Q3: $\chi^2(1, n = 115) = 0.89$, $p = .35$; Q4: $\chi^2(1, n = 115) = 0.00$, $p = .98$; total number of correct answers: $\text{mean}_{\text{think aloud}} = 2.20$, standard $\text{deviation}_{\text{think aloud}} = 0.76$; $\text{mean}_{\text{written}} = 2.25$, standard $\text{deviation}_{\text{written}} = 0.83$; $t(113) = 0.31$, $p = .76$).

Syst. Res. **31**, 268–279 (2014)
DOI: 10.1002/sres

Table 2 Results stock-flow tasks (N = 115)

Answers	Q1 most entering		Q2 most leaving		Q3 most in stock		Q4 fewest in stock	
	n	%	*n*	%	*n*	%	*n*	%
Max entering $t = 4$	**112**	**97.4**	4	3.5	5	4.3		
Max leaving $t = 21$	1	0.9	**105**	**91.3**	1	0.9	6	5.2
Max in stock $t = 13$					**24**	**20.9**	8	7.0
Fewest in stock $t = 30$					1	0.9	**16**	**13.9**
Max net inflow $t = 8$			2	1.7	67	58.3		
Max net outflow $t = 17$					5	4.3	65	56.5
Initial in store $t = 1$	1	0.9	1	0.9			3	2.6
Cannot be determined					12	10.4	17	14.8
Other	1	0.9	3	2.6				
No answer								

Combination of experimental conditions, department store and bank task, and think aloud and written condition. 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 min of the correct response). Conforming Cronin *et al*. (2009, p. 119), bold numbers indicate correct responses, and underlined numbers show the incorrect answers that give the maximum net inflow or net outflow (indicated by the largest difference between inflow and outflow in Figure 1) instead of maximum/minimum in the stock.

The interaction effect between mode and task context on Total number of correct answers was also not statistically significant ($F < 1$). Therefore, we combined the numerical results of both experimental conditions (Table 2).

Table 2 shows comparable patterns of task-flow performance as reported in previous research (e.g. Sterman, 2002; Pala and Vennix, 2005; Cronin *et al*., 2009). Participants generally do not have any problems answering Q1 and Q2. The limited number of correct answers of Q3 and Q4 indicates that participants in this study also had difficulties with the concept of accumulation. In addition, the high percentages in the underlined cells in column 'Q3 most in stock' and column 'Q4 fewest in stock' reveal that most of the participants incorrectly focus on the maximum net inflow or outflow instead of on the maximum or minimum of the stock.

The distribution of the total number of correct answers in Table 3 shows that 12.2% of the participants completed the task without error.

The transcription and analysis of the video tapes of 44 participants with a focus on the reasoning used for answering Q3 and Q4 resulted in reasoning types and claims of the participants (Table 4). The first column of Table 4 shows the description of the reasoning lines A through M as well as their categorization in the seven reasoning types (1 through 7). Each reasoning line is illustrated with a text fragment in column two (printed in italics). When the argumentation resulted in a correct claim, the participant number was placed in column 'Claim Q3 +' and/or in column 'Claim Q4 +', whereas in the case of incorrect claims, the participant number was positioned in column 'Claim Q3 −'and/or in column 'Claim Q4 −'. For example, participant number 7 reasoned in numbers only, followed by incorrect claims for both Q3 and Q4.

Table 3 Frequency of total number of correct answers (N = 115)

	n	%
0	1	0.9
1	9	7.8
2	82	71.3
3	9	7.8
4	14	12.2

The results of the think aloud analyses presented in Table 4 reveal that a limited number of participants gave a correct argumentation followed by correct claims (reasoning line M; Q3: 5 (11%); Q4: 5 (11%)) and that in a few instances, ambiguous or absence of explicit argumentation resulted in a correct claim (reasoning lines B, J and K; Q3: 3 (7%); Q4: 2 (5%)).

The think aloud method enabled us to explore the significance of correlational reasoning in situations where participants made an incorrect claim.

Syst. Res. **31**, 268–279 (2014)
DOI: 10.1002/sres

Table 4 Think aloud: reasoning types and lines, illustrations and claims, question 3 (Q3) and question 4 (Q4; n = 44)

Reasoning type (1–7)	Illustration	Claim Q3[a]		Claim Q4[a]	
Reasoning line (A–M)		+	−	+	−
1. Absence of explicit reasoning					
A. 'Think aloud', only in numbers.	'higher than 47, 30, 27, 37 minus 25, 12 (. . .) lower, so 16 (. . .)'		7		7
B. Claim straight away.	'and the fewest (. . .) after 30 minutes'		20	*21*	*17* 20
2. Flow: either inflow or outflow					
C. Highest peak inflow or highest peak outflow.	'because then the withdrawals are highest so you have the fewest Euros in the bank'		36		36
3. Inflow–outflow difference at a specific point in time					
D. Minimum of the difference between inflow and outflow at a specific point in time.	'fewest (. . .) when the difference is smallest, I guess'				*33* *35*
E. Maximum of the difference between inflow and outflow, at a specific point in time.	'the biggest difference between entering and leaving'		1 4 5 6 8 9 10 13 *14* 15 *17* 19 22 *23* 24 28 *29* 30 31 *33* *35* 37 38 40 42 43		1 4 5 6 8 9 10 13 15 19 22 24 28 30 31 37 38 40 42 43
4. Stock-stock relation					
F. Unable or difficult to determine, the stocks, because of a lack of information about the absolute numbers.	'I just think that you can't determine the exact point in time, because you have no exact numbers'		3		3 *14* *32*
G. Difficult to determine the exact relation between accumulation periods.	'because it is a relative decrease of the amount of money in the bank, but because of the period before this [decrease] could be undone'				*27*
5. Reasoning mix					
H. The average of a number of points in time, in addition to the determination of the highest peak (inflow).	'taking the average of minutes, may be (. . .), however, (. . .) I still go for minute four'		18		18
I. Maximum of the difference between inflow and outflow at a specific	'that would be the difference between entering and leaving at that point (. . .) I'm not sure		25 *32* 39 41 44		25 39 41 44

DOI: 10.1002/sres

point in time in addition to a lack of information about absolute numbers.	(. . .) you don't know how many people were already inside'			
J. Accumulation in addition to a lack of information about absolute numbers.	'still more entered than withdrawn (. . .) I think you just can't determine the exact point (. . .) the higher the peak the more Euros are entered (. . .)'	*16* 21	*26*	*29*
6. Inflow–outflow relation: process and accumulation?				
K. Intersection of inflow and outflow.	'most people in (. . .), evidently, you have to take an intersection point'	26		
7. Inflow–outflow relation: accumulation				
L. Determination of the minimum as the accumulation of negative net flow during part of the trajectory.	'from minute 8 until 17 or 18, an increasing number of people are leaving and in the same period less and less people are entering'			23
M. Determination of the maximum/minimum as the accumulation of the (positive/negative) net flows.	'because till then more was deposited than withdrawn'	2 11 12 *27* 34	2 11 12 *16* 34	

Claim: +, correct; −, incorrect.
[a]Cells contain participant numbers, numbers underlined and italicised show that the participant used different reasoning *lines* for Q3 and Q4.

Table 4 shows that most of the participants who came up with an incorrect claim used a line of argument in accordance with the stock-(net) flow correlation (reasoning lines C, D and E; Q3: 27 (61%); Q4: 23 (52%)). The majority of these participants used a stock-net flow line of reasoning, by mistakenly relating the difference between inflow and outflow to the maximal/minimal value of the stock (reasoning line E; Q3: 26; Q4: 20). Participant 28 verbalized his reasoning as: 'the biggest difference between entering and leaving'. Only in a few instances a stock-flow correlation was used in which the highest peak of the inflow/outflow is taken as the point in time where the stock is maximal/minimal (reasoning lines C and D; Q3: 1; Q4: 3). Hence, the participants' verbalization allowed us to differentiate between stock-flow and stock-net flow correlation heuristic with the latter being the more dominant (cf. Sterman, 2010, p. 328).

In addition, the think aloud analyses show that besides the correlation heuristic, there is a number of other reasoning lines that led to incorrect claims (Table 4, reasoning lines A, B, F, G, H, I, J and L; Q3: 9 (20%); Q4: 14 (32%)). Of particular interest is the outcome that a number of participants, although well aware of the stock character of Q3 and Q4, erroneously assumed that they needed to have information about the absolute numbers at $t = 0$ (reasoning lines F, I, and J; Q3: 6; Q4: 8). Some participants were very explicit in their request for absolute numbers:

> 'You can't determine that either, because you don't know how many, you do know how many Euros were deposited per minute, but you don't know what the beginning is, so it can't be determined' (participant 32).

Whereas others were less secure: 'I'm not sure (. . .) you don't know how many people were already inside' (participant 25).

Interestingly, Table 4 (underlined numbers) also shows that 11 participants use different reasoning lines for solving Q3 and Q4, respectively. From a learning perspective, some participants are of particular interest. Participant 21 tackled Q3 from different perspectives:

> 'The higher the peak, the more Euros are added per minute, so here 40 are added per minute and 20 are taken off 40. So then there are less Euros in the bank. Oh you obviously can . . . That would be this (points to graph). Oh I actually think something else. Because here there's an increase in Euros per minute. Yes, obviously also minute 13. And the fewest . . . In minute hmm 30. Now, this is the way I think it is (. . .)'.

Apparently, this analytical zigzagging needed to solve Q3, gave the participant the insight how to handle this type of problem, for she gave the—correct—answer at Q4 in just a few seconds, without deliberating. A similar approach is taken by participant 16. He also approached Q3 from the various perspectives mentioned in the previous texts, and eventually, he too came up with the correct answer. The trial and error method used to solve Q3 also led to a better understanding of the precise nature of Q4, although less explicit as compared with participant 21.

Apart from the cognitively oriented reasoning lines discussed earlier, several participants appeared to have interpretation problems related to information presented in the stock-flow task (Table 5).

The interpretation problems in Table 5 refer to the terminology used (I) as well as to the presentation of the graph (II and III). Although we deemed a bank task appropriate for the Business administration students in our sample, two participants had problems with terminology used. With respect to category II, the participant assumed that the dissimilar ways the inflow-lines and outflow-lines were portrayed represented different meanings: the unbroken, bold inflow-line was believed to stand for hard data, whereas the data behind the dotted outflow-line were assumed to be more tentative. Four participants mentioned interpretation problems with the *y*-axis of the graph (category III). In particular, they had difficulties with the labelling of the *y*-axis as a ratio. They expected the label 'people' or 'Euros', and not 'people *per* minute' or 'Euros *per* minute'. This observation is of importance, because up to now, the discussion about the ability of participants to interpret graphs focuses on differences in format (line graphs versus tabular format, bar graphs

Syst. Res. **31**, 268–279 (2014)
DOI: 10.1002/sres

Table 5 Interpretation problems in think aloud condition, question 3 (Q3) and question 4 (Q4)

Interpretation problems	Illustration	Claim Q3[a]		Claim Q4[a]	
		+	−	+	−
I. Unfamiliar with bank terminology.	'sorry, deposited means that (. . .)?'		15 35		15 35
II. Ambiguity regarding the layout of the graph lines.	'what is striking is that indeed the entering line is in bold, so I think that that is perhaps certain, while the leaving line is a dotted line, so I can imagine that that is an estimation'		18		18
III. Ambiguity regarding *y*-axis labelling, use of a ratio containing a slash (/).	'but in that case the left column should be just people but it is people per minute'		5 6 8 24		5 6 8 24

Claim: +, correct; −, incorrect.
[a]Cells contain participant numbers.

and text; Cronin *et al*., 2009) and not on more specific features of the graph itself. Locating the flow as a ratio on the *y*-axis may be a convention in the field of system dynamics, but in other disciplines, a ratio is obtained by combining the absolute numbers of the *y*-axis with the time location on the *x*-axis, for example, in the field of economics, a graph of the development of Gross National Product (*y*-axis) per year (*x*-axis). The influence of these, rather specific, interpretation problems did not receive attention in the literature so far.

CONCLUSION AND DISCUSSION

Our study confirms existing stock-flow research instead of literature in several aspects. We studied reasoning patterns as verbalized by participants performing stock-flow tasks. In an experimental 2×2 factorial design, participants were randomly assigned to a task, department store or bank, and a verbalization mode, verbal (think aloud) or written. The different tasks yielded no differences in performance, measured as the correct answer to the four questions and the total number of correct questions, indicating that there was no evidence for a task context effect. This finding is in line with previous research (e.g. Cronin *et al*., 2009; Brunstein *et al*., 2010). There was also no effect of the factor mode: participants in the verbal and written condition showed no performance differences. Apparently, verbalization in comparison with writing does not improve or hinder task performance. In line with Fox *et al*. (2011), we find that verbalization is nonreactive: this mode of data collection did not yield different outcomes than the written mode. We tentatively conclude that the experimental manipulations did not influence the outcomes. The combined task results of the current study appeared comparable with previous research. Only 12% of the participants completed the task without error. Furthermore, from the large number of participants incorrectly interpreting questions in terms of max net inflow/outflow (Q3 and Q4, respectively), we infer that many participants in the current study also had problems with the concept of accumulation (cf. Sterman, 2010). In addition, our findings contribute to the existing scientific evidence pointing to the absence of a task context effect in stock-flow performance (Cronin *et al*., 2009; Brunstein *et al*., 2010). Individuals' achievements appear to be independent of the domain or task context. The activation and application of stock-flow knowledge seem not to be influenced by the salience and familiarity of the accumulation process in the task context.

Our research makes a number of novel contributions to the literature as well. This study is one of the first to apply the think aloud method to stock-flow tasks. Using this method enables us to examine the reasoning patterns of participants while they perform stock-flow tasks. This resulted in several outcomes. First, the method enabled us to corroborate the correlation heuristic. So far,

Syst. Res. **31**, 268–279 (2014)
DOI: 10.1002/sres

research outcomes strongly supported the existence of this heuristic, but the factual reasoning processes were not examined. In our study, many participants verbalized their thoughts in terms of the biggest difference between inflow and outflow at a particular point in time, thus expressing the correlation heuristic in words.

Second, we also discovered that participants approach the stock-flow task in different ways, varying from the absence of any explicit reasoning, through a use of a reasoning mix, to an approach based on a comprehensive understanding of the accumulation process (Table 4). So besides the dominant approach in the form of the correlation heuristic, some participants in our sample also use other reasoning strategies, which led to incorrect claims.

A third outcome the think aloud method revealed was the perceived lack of an initial value of the stock. Although the initial stock value is unnecessary to answer Q3 and Q4 correctly (Sterman, 2002), about one-fifth of the participants wanted more information about absolute numbers. This apparent contradiction might be caused by the fact that the department store (or bank) task shows little resemblance with real-life problem solving. Normally, the search for options or solutions starts when there is a 'mismatch between outcome and expectation' (Argyris and Schön, 1978, p. 18). The results of former actions or the actual state of a particular phenomenon are compared with a reference point, norm or standard to decide on the problematic character of a situation and on the need for corrective actions. Considered on its own, a temporary excess of Euros withdrawn over Euros deposited is not problematic for a bank. To decide in this matter, information is needed about the initial value, in this case, about the amount of money already in the bank. So, taking into account initial numbers is a rather common strategy in diagnosing the problematic character of a situation. The absence of an initial value may have hampered the execution of the task. The influence of information on the initial value of the stock could quite easily be studied in an experiment using the classical task format as a control condition.

The fourth outcome we discussed is interpretation problems with the terminology of the *y*-axis of the graph as a ratio (Table 5). Apparently, for some participants, elements of the task, such as the format of the graph, distracted them from the solution of the actual task. Thus, in further experimental research, the effect of labelling of the *y*-axis on task performance should be tested.

The final and perhaps most intriguing outcome of this study is that in dealing with stock-flow problems, participants switched between reasoning lines, as well as between reasoning lines and interpretation problems. From a learning perspective, these switches are of special interest. They apparently represent a process of 'trial and error' that together with a 'mechanism of comparison' is conceived as essential in learning processes (Bateson, 1972. pp. 248, 287). By feeding back on errors, we are not only able to 'solve particular problems but also form *habits* which we apply to the solution of *classes* of problems' (Bateson, p. 274, italics in original). Once the solution to a problem is found, this expertise is stored in memory. Faced with a similar problem, the correct, intuitive response to this problem is 'nothing more and nothing less than recognition'. Analogous problems can then be understood at a glance, the person concerned knows intuitively, 'unable to describe in detail the reasoning or other process that produced the answer' (Simon, 1992, p. 155). So, to move to a higher level of learning or problem solving, it is essential to decrease the ease of heuristic reasoning and to trigger a process of trial and error. A possibility to evoke such a mental process of trial and error in the department store task is to ask participants if their answers to Q3 and Q4 represent their first impulse or subsequent, more deliberate, reasoning strategies. This approach may provoke a process in which alternatives are mentally simulated.

In the literature, it is often stated that the stock-flow problems presented are simple, but we experienced that this did not hold for the participants in our study. In line with this, we also observed that emotional aspects play a role. Individuals do not like to fail and lose face. This is apparent from verbalizations such as 'I think it should be very easy to answer these questions, but I am making a mess of it', '... hmm ... this is very difficult', and 'I feel somewhat like an idiot'. So, besides trying to increase our insight in reasoning strategies,

Syst. Res. **31**, 268–279 (2014)
DOI: 10.1002/sres

research may also pay attention to face-saving aspects related to task performance. This means that not only the cognitive aspects are important but also emotional reactions and coping behavior to manage an uncomfortable situation. It is thus important to focus on emotional competencies like self-awareness, self-management and social awareness, and their possible relationship to stock-flow task performance in a secondary analysis of the think aloud method (cf. Goleman, 2000).

REFERENCES

Argyris C, Schön D. 1978. Organizational Learning: A Theory of Action Perspective. Addison-Wesley: Reading, MA.

Bateson G. 1972. Steps to an Ecology of Mind. Chandler: San Francisco.

Boeije H. 2010. Analysis in Qualitative Research. Sage: London.

Booth Sweeney L, Sterman JD. 2000. Bathtub dynamics: initial results of a systems thinking inventory. *System Dynamics Review* **16**: 249–286.

Brunstein A Gonzalez C, Kanter S. 2010. Effects of domain experience in the stock-flow failure. *System Dynamics Review* **26**: 347–354.

Cohen J. 1992. A power primer. *Psychological Bulletin* **112**: 155–159.

Cronin M, Gonzalez C, Sterman JD. 2009. Why don't well-educated adults understand accumulation? A challenge to researchers, educators, and citizens. *Organizational Behavior and Human Decision Processes* **108**: 116–130.

Ericsson KA, Simon HA. 1993. Protocol Analysis: Verbal Reports as Data (revised edn). MIT Press: Cambridge, MA.

Fox MC, Ericsson KA, Best R. 2011. Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin* **137**: 316–344.

Goleman D. 2000. Leadership that gets results. *Harvard Business Review* **78**: 78–90.

Kahneman D, Slovic P, Tversky A. (eds.). 1982. Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press: Cambridge.

Kainz D, Ossimitz G. 2002. Can students learn stock–flow-thinking? An empirical investigation. *Proceedings of the 20th International Conference of the System Dynamics Society, Palermo, Italy.* System Dynamics Society: Albany, NY.

Ossimitz G. 2002. Stock–flow-thinking and reading stock-flow-related graphs: an empirical investigation in dynamic thinking abilities. *Proceedings of the 20th International Conference of the System Dynamics Society,* Palermo, Italy. System Dynamics Society: Albany, NY.

Pala O, Vennix JAM. 2005. Effect of system dynamics education on systems thinking inventory task performance. *System Dynamics Review* **21**: 147–172.

Phuah TLK. 2010. Can individual learn behaviours of stock and flow using their ability to calculate running total? An experimental study. *Proceedings System Dynamics Conference*, Seoul, Korea.

Simon HA. 1992. What is an explanation of behavior? *Psychological Science* **3**: 150–161.

van Someren MW, Barnard YF, Sandberg JAC. 1994. The Think Aloud Method. A Practical Guide to Modeling Cognitive Processes. Academic Press: London etc.

Sterman JD. 1989. Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment. *Management Science* **35**: 321–339.

Sterman JD. 2002. All models are wrong: reflections on becoming a systems scientist. *System Dynamics Review* **18**: 501–531.

Sterman JD. 2010. Does formal system dynamics training improve individual's understanding of accumulation? *System Dynamics Review* **26**: 316–334.

Strauss AL, Corbin J. 2007. Basics of Qualitative Research. Techniques and Procedures for Developing Grounded Theory (3rd edn). Sage: Thousand Oaks, CA.

Tversky A, Kahneman D. 1974. Judgment under uncertainty: heuristics and biases. *Science* **185**: 1124–1131.

Syst. Res. **31**, 268–279 (2014)
DOI: 10.1002/sres