Paper presented at the 33rd International Conference of the System Dynamics Society,
July 19-23, 2015, Cambridge, MA

Bathtub Dynamics Revisited: Exploring the engineering domain

Florian Kapmeier[1*], Meike Tilebein[2], and Roland Maximilian Happach[2],

[1] ESB Business School, Reutlingen University
Alteburgstraße 150, 72762 Reutlingen, Germany

[2] Institute for Diversity Studies in Engineering, University of Stuttgart
Pfaffenwaldring 9, 70569 Stuttgart, Germany

*corresponding author: florian.kapmeier@reutlingen-university.de

Version 1.1

August 2015

Abstract

SF-failure, the inability of people to correctly determine the behavior of simple stock and flow structures is subject of a long research stream. Reasons for SF-failure can be attributed to different reasons, one of them being lacking domain specific experience, thus familiarity with the problem context. In this article we present a continuation of an experiment to examine the role of educational background in SF-performance. We base the question set on the Bathtub Dynamics tasks introduced by Booth Sweeney and Sterman (2000) and vary the cover stories. In this paper we describe how we developed and tested a new cover story for the engineering domain and implemented the recommendations from a prior study. We test three sets of questions with engineering students which enables us to compare the results to a previous study in which we tested the questions with business students. Results mainly support our hypothesis that context familiarity increases SF-performance. With our findings we further develop the methodology of the research on SF-failure.

Keywords: Bathtub dynamics, stock-and-flow-performance, context familiarity, domain specific knowledge, educational background

Introduction

SF-failure, the inability of many individuals to correctly deal with even simple systems' stock and flow-(SF-)structures and their behaviors, is a widely discussed topic that has been intriguing scholars for almost two decades. Building on the survey method - a written test with a number of different tasks - and the insights from the groundbreaking original study by Booth Sweeny and Sterman (2000), many scholars have since contributed additional perspectives and details to shed light on the phenomenon of SF-failure (Kapmeier & Zahn, 2001; Kainz & Ossimitz, 2002; Ossimitz, 2002; Sterman, 2002; Kapmeier, 2004; M. A. Cronin & Gonzalez, 2007; M.A. Cronin, Gonzalez, & Sterman, 2009; Moxnes & Jensen, 2009; Brunstein, Gonzalez, & Kanter, 2010; Kapmeier, Happach, & Tilebein, 2014; Fischer & Gonzales, in press).

SF-failure is a relevant phenomenon as people have to deal both in professional and in everyday life contexts with ever more complex issues. The political, ecological, social, technological, legal and economic environments that surround decision-makers are interrelated and full of delays, nonlinearities and SF-structures. If we knew more about the phenomenon of SF-failure and how to mitigate it, people could be better prepared to deal with such complex issues in an appropriate way.

Influences on SF-performance are found in three fields: 1) systems thinking, 2) visualization, and 3) domain-specific knowledge and experience (Kapmeier et al., 2014). Within these fields, however, studies reveal contrary findings, effects are debated, and still many topics have to be further explored in order to finally fully explain and thus improve SF-performance

First, in the field of general systems thinking aspects, an effect that was found to be prevalent in many previous studies (Sterman & Booth Sweeney, 2002; M.A. Cronin et al., 2009; Sterman, 2010; Korzilius, Raaijmakers, Rouwette, & Vennix, t.b.a.) is termed pattern matching (Booth Sweeney & Sterman, 2000).This means that individuals, when asked to predict the output behavior of an SF-system, intuitively match the output exactly with a system's input. As an illustrative example, Sterman (2002) describes the violation of conservation of matter when matching a desired decline of CO_2 accumulation to an according (and false) pattern of inflow of CO_2 emissions .

Second, in the field of visualization effects, scholars look at differences in presentation of SF-tasks and their correlation with SF-performance. Instead of graphs, like in the initial survey, their studies use other forms of illustrations (M.A. Cronin et al., 2009; Schwarz, Epperlein, Brockhaus, & Sedlmeier, 2013; Sedlmeier, Brockhaus, & Schwarz, 2014), lab experiments (Größler & Strohhecker, 2012), or tables (Kainz & Ossimitz, 2002).

Third, investigating effects of domain-specific knowledge and experience, scholars hypothesize that context familiarity plays a role in SF-performance. In particular, unfamiliarity with a SF-task's context contributes to SF-failure, and likewise participants perform better in tasks related to their specific knowledge domain. Previous studies mainly used educational background, e.g. field of study, majors, or

credits in specific courses including beer distribution game as proxies for domain-specific experience and either tried to exclude the effect of domain-specific experience by designing tasks from an everyday life context (e.g. (Booth Sweeney & Sterman, 2000)) or designed tasks to examine this effect (e.g. (Brunstein et al., 2010)). The results of our previous study (Kapmeier et al., 2014) indicate that, aside of systems thinking skills, prior domain-specific knowledge may have a larger influence on SF-performance than expected.

The three fields of effects on SF-performance are not independent. As an example, pattern matching is a common error subjects might be especially prone to if graphs are means of problem representation. In turn, the question of problem representation is closely related to the question of context familiarity, as members of a specific discipline typically share the discipline's specific culture, including preferences towards means of representation and visualization. In addition, representation issues imply a multi-level perspective, as individual preferences add to the shared preferences from a discipline's 'common ground' of visualization. Thus clarifying the effect of domain-specific experience on SF-performance could bear additional insights also for the other two fields affecting SF-performance. Moreover, we assume that clarification of this effect might be especially interesting for determining whether education in dealing with SF-systems should rather be part of curricula on the level of domain-specific knowledge instead of being taught on the level of general skills and independent from specific knowledge domains.

In a first study on this topic (Kapmeier et al., 2014) we designed two additional cover stories in order to adapt the respective square wave task of Booth Sweeney and Sterman's (2000) survey to the knowledge domains of management and engineering, respectively. Each of the new cover stories had the same underlying SF-structures and dynamics as the task in the original study. In our study we presented the newly framed square wave tasks, as well as the original one addressing a daily life context, followed by the original sawtooth task, to three groups of business students. Findings suggest that familiarity with the context, thus domain-specific knowledge and experience, does have a large impact on SF-performance: Business students performed much better if the task was related to a business context than if the task was from a different knowledge domain. Interestingly, students even performed better on a more challenging SF-structure (the sawtooth task) from their own knowledge domain than on a simpler one from a more distant domain.

Yet, in thoroughly analyzing and discussing the results we identified a number of potential biases regarding method and procedure. In addition to eliminating these we now aim at enhancing statistical relevance of our findings by including more participants. With this article, we present the results of additional studies we conducted in order to address a number of the biases. In particular, we explore other sources of domain-specific experience by expanding the cover sheet, we re-design the cover story addressing the engineering domain, we ensure task sequence and completion time, and we complement our previous findings with participants with educational background in the engineering domain. While

some of the new results confirm our previous expectations, others contribute to the more general debates on testing systems thinking skills on the one hand, and on teaching systems thinking skills on the other hand, particularly whether this should take place with a domain-specific or rather with a general focus.

The research question of this article translates into the following hypothesis: SF-performance increases when the problem context is embedded in the educational background of the problem solver. This hypothesis was confirmed with participants from a business study program in our first study (Kapmeier et al., 2014). With our new study we gather additional data from the engineering knowledge domain that also supports our hypothesis.

This article is organized as follows. In the next section, we present a short recapitalization of insights from our previous study, as well as remaining open questions. Afterwards, we state how we addressed the open questions in our new study, eliminating potential biases we had identified in our previous study regarding participants, procedure, and method. Then, we describe the steps of the implementation process, along with its intermediate and final results, respectively. We then discuss the new results both individually and against the results of our previous study. Finally, we discuss limitations and implications as well as future research paths in the closing section.

Previous insights and new study

Based on the stream of literature indicating that domain-specific knowledge and experience as well as familiarity with the problem context may play a role in SF-performance, we set up a study to further investigate this issue. In line with previous studies we took educational background and field of study in particular as indicator for domain-specific experience. We included three groups of participants, i.e. a control group who received exactly the original study's tasks, and two Experimental Groups. In the task sheets for the Experimental Groups, we changed the original study's bathtub cover story, meant to address an everyday life situation that is familiar to most individuals and not related to a specific knowledge domain, into cover stories tailored to the knowledge domains of business and engineering, respectively. To this first task with underlying square wave dynamics we added the cash flow task with underlying sawtooth dynamics from the original study, describing a company's cash balance (i.e. the stock) that is changed by receipts (i.e. the inflow) and expenditures (i.e. the outflow). The identical task sequence and underlying behavioral patterns in the test sheets for the three groups are shown in Figure 1.

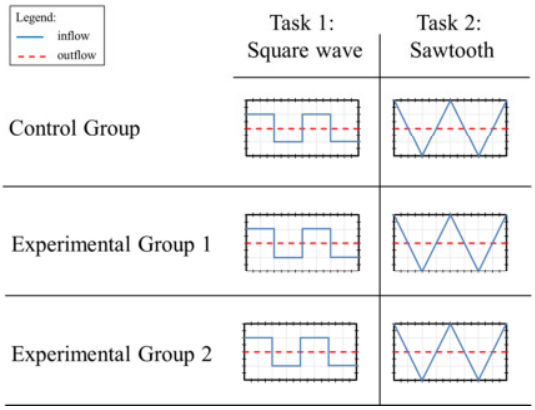


Figure 1: The sequence of behavioral patterns for inflow (solid blue line) and outflow (dotted red line) is the same for all three sets of questions. Source: (Kapmeier et al., 2014)

Figure 2 shows the similar visualization of the three alternative square wave tasks. A sketch of a bathtub (i.e. the stock) with water inflow on the left and outflow on the right comes with the original bathtub cover story. Similarly, the Application cover story for Experimental Group 1 describes a situation in which online job applications (a pile of files) are the stock that is filled by an inflow (a 'send e-mail' icon with an '@'-key) from the left, while applications are withdrawn, visualized by a 'delete' key and a recycling bin icon above the outflow at the right hand side. This task is topic of human resource management and therefore relates to the business domain. In a similar way, the engineering-related Vehicle cover story is visualized by a car's speedometer with a race car (acceleration) as inflow and braking lights shining brightly (braking) as an outflow on the right. As mentioned above, the three tests shared the same second task.

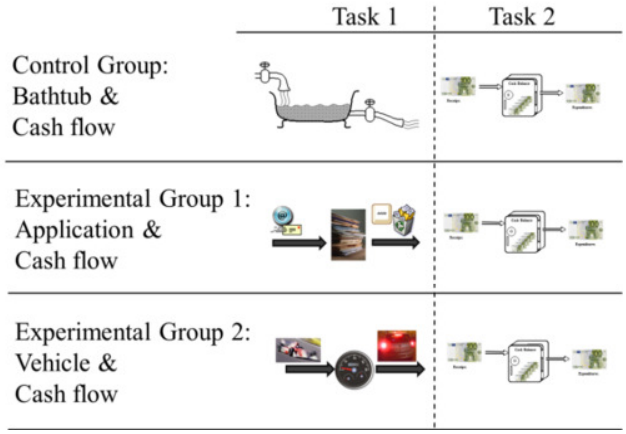


Figure 2: The cover stories: Bathtub & Cash flow, Application & Cash flow, and Vehicle & Cash flow with their respective visualizations for the sets of questions containing task 1 and task 2 (stocks with their inflows and outflows). Source: (Kapmeier et al., 2014)

With this experimental setup we aimed at evaluating our hypotheses: Stock and flow performance increases when the problem context is embedded in the educational background of the problem solver. Participants in our previous study were business students. We used the Control Group's results derived from the original tasks from Booth Sweeney and Sterman (2000) to compare and align our study to the outcomes of the original and subsequent studies.

Regarding our test described above, we expected that domain-specific tasks with simple behavioral patterns (i.e. square wave pattern) from the problem solver's knowledge domain are better solved than tasks that are from a more distant knowledge domain. Hence we expected that in a study with business students, the Experimental Group 1 with the Application cover story would cope best with the first task, and that Experimental Group 2 would cope worst. Further, we expected that the three groups would show somewhat equal performance on the second task that was identical in the three groups.

Moreover, we wondered if the effect of domain-specific experience could even overrule the effect of higher task difficulty to a certain extent, e.g. subjects might perform better in a more complicated systems thinking task (i.e. sawtooth pattern) which is embedded in their educational background than in a simpler task of which the problem context is rather distant from their educational background. Hence we suggested that business students from Experimental Group 2, even if they failed to solve the Vehicle cover story, might show the same performance as the Control Group and the Experimental Group 1. Regarding the expectations formulated above, our study revealed the following results as shown in Table 1.

		confirmed	supported	not supported	contradicted	biased?
The Experiment	Overall hypothesis at large		x			x
	Comparable cover stories		x			x
	Alignment with original study		x			
	Alignment with other studies		x			
Results of the Business Students	Overall performance in Application cover story better than in Vehicle cover story		x			
	Overall performance in Application cover story better than in original Bathtub cover story		x			
	Overall performance in Bathtub cover story better than in Vehicle cover story		x			
	Relatively better performance in complicated task that is business related than in easy task from distant knowledge domain		x			x
	SF-performance (avoiding pattern matching) is much better in Application vs. Bathtub vs Vehicle cover stories	X				x
	Sawtooth task is more difficult, just like in previous studies, leading to lower average performance than in square wave task				x	x
	All three groups, as they have the same background, perform equally well on sawtooth task (second task), independent from first task			x		x
	Pattern matching is most common error in sawtooth task, just like in previous studies	X				

Table 1: Expectations and findings from our previous study with business students (Kapmeier et al., 2014)

Discussing these results, we pointed to three areas of limitations and potential biases in the evaluation of our hypotheses, these are participants, procedure, and method. In particular, we raised a number of issues which we address in our new study by the following measures, listed by area:

1. *Participants*
   a. Increase the overall number of participants: In our first study we gathered data from about 40 students grouped into two Experimental Groups and one Control Group, to test the two new domain-specific cover stories for the square wave task in parallel to the original cover story. Now we complement the material with data sets from additional 50+ students.
   b. Extend the range of knowledge domains throughout participants: While in our first study we approached business students from the BSc program at ESB Business School Reutlingen, we now turn to comparable engineering students in order to mirror our findings from the first study. We carry out the new study with first semester students and higher semester students enrolled in a

Mechanical Engineering program (BSc Engineering Cybernetics) at the University of Stuttgart, Germany.

2. *Procedure*
    a. Control the maximum time spent on each of the two tasks during the test: By detailing the participants' instructions preceding the test, implementing a new procedure to make sure that participants start working on the sawtooth task only after having spent 5 min working on the square wave task, and controlling for strict compliance to that new procedure, we enforce an exact time split with participants working for 5 min on each task.
    b. Control the sequence of work during the test: The new procedure used to ensure the time split also guarantees that participants work on the first task (square wave) first.
3. *Method*
    a. Ensure the comparability of cover stories from different knowledge domains: In our previous study, the new cover stories, particularly the one addressing the engineering domain, differ from the original cover story in terms of tangibility and countability. Further, we used an engineering-specific task with rather general descriptions. The terms 'braking' and 'acceleration' might cause confusion. We believe that the terms might be too general for engineers but also represent lingual problem since these two terms are mutually exclusive in German. We address this potential bias by adjusting the engineering-related cover story for the square wave pattern to match the original task with respect to these two dimensions.
    b. Extend the assessment of prior domain specific experience: For our new study, we integrate additional questions to the respective section of the background data sheet in order to detect potential domain specific experience from additional sources the original background data sheet does not capture, e.g. prior work experience (Booth Sweeney & Sterman, 2000).

Implementation Process and Results

In order to continue and deepen our research on systems thinking skills and educational background we started with updating the cover sheet (measure 3b). In addition to the existing questions (Booth Sweeney & Sterman, 2000) we asked the participants to provide additional information on vocational training and internships they might have done. As can be seen in the appendix, participants are also asked to specify the area in which they had gathered this previous working experience.

We also developed a more rigorous process to track the time to work on the questionnaire (measures 2a and 2b). The coversheet had already included many questions that usually take the participants some time to fill out. As in previous studies, we waited with the beginning to work on the first question until everybody was done filling out the cover sheet. Participants are informed that they have 5

minutes to work on each question. They are being told that they are not allowed to turn to the second question before 5 minutes were up. The moderator tells the participants to turn to the second question exactly after 5 minutes – a stop watch helps to track the time. Participants then have 5 minutes to work on the second question. Yet, it is not strictly prohibited to revisit the first question.

As indicated above, one further objective of our study is to enlarge group of participants to a group of students comparable to the ones in our previous study (Kapmeier et al., 2014). We thus conducted the questionnaire with twenty-two 5th semester students of the BSc Engineering Cybernetics and first semester MSc Engineering Cybernetics study programs of the University of Stuttgart in November 2014 (measure 1b). This was also the first group with which we tested the new coversheet and the more thorough approach to track the time for each question.

As can be seen from the Table 2, the background data sheet asked for information about the participants' age, gender, and current degree program, vocational training, internship, region of origin, first language, highest previous degree, teaching language, and whether they had played the beer distribution game before.

The proportion of male and female participants is unevenly distributed in all groups with a majority of male students (68%). All but one (5%) of the participants were enrolled in either the BSc (50%) or the MSc (45%) program in Engineering Cybernetics at the University of Stuttgart. This individual was enrolled in the MSc in Technology Management program. All participants were younger than 24 years old and the great majority is from Europe, more precisely, from Germany (91%). Two participants originate from outside Europe (9%), one from Asia and one from the Middle East. German, the language of the study programs, was the first language for the great majority of participants (90%). None of the participants' first language was English, the language of the task. Fifty percent of the participants had a BSc or equivalent degree, all of them in the area of engineering. Nobody had previously done a vocational training and only two participants (9%) had done an internship in engineering. None of the participants had played the beer distribution game before.

	University of Stuttgart Winter 2014/2015 Engineering Cybernetics November 2014			
Task	**Total numbers**	**Control Group: Bathtub and Cash Flow**	**Experimental Group 1: Application and Cash Flow**	**Experimental Group 2: Vehicle and Cash Flow**
Total Number of Students	**22**	**8**	**7**	**7**
Age				
19-24	22	8	7	7
25-30	0	0	0	0
31-35	0	0	0	0
36 and up	0	0	0	0
Gender				
Male	15	4	5	6
Female	7	4	2	1
Student Status				
Engineering Cybernetics	21	8	7	6
Technology Management	1	0	0	1
Prior Field of Study				
Business/Management	0	0	0	0
Engineering	11	3	4	4
Social Science	0	0	0	0
Computer Science	0	0	0	0
Mathematics	0	0	0	0
Humanities	0	0	0	0
Highest Prior Degree				
BA	0	0	0	0
BSc	10	3	4	3
MA, MSc, Diplom	0	0	0	0
Ph.D.	0	0	0	0
High School	11	5	3	3
BE. JD, BBA, MD, CPA	1	0	0	1
Current Field of Study				
Business/Management	0	0	0	0
Engineering	22	8	7	7
Social Science	0	0	0	0
Science	0	0	0	0
Computer Science	0	0	0	0
Mathematics	0	0	0	0
Humanities	0	0	0	0
Vocational training				
Business/Management	0	0	0	0
Engineering	0	0	0	0
IT	0	0	0	0
Other	0	0	0	0
No vocational training	22	8	7	7
Internship (=> 3 months)				
Business/Management	0	0	0	0
Engineering	2	0	0	2
IT	0	0	0	0
Other	0	0	0	0
No vocational training	20	8	7	5
Region of Origin				
North America (Aus. + NZ)	0	0	0	0
Europe	20	7	7	6
Asia and Middle East	1	0	0	1
Latin and South America	0	0	0	0
Africa	1	1	0	0
First language				
German	20	7	6	7
English	0	0	0	0
Other	2	1	1	0
Teaching Language				
English	0	0	0	0
German	22	8	7	7
Beer Game Experience				
Played before	0	0	0	0
Have not played	22	8	7	7

Table 2: Subject demographics – University of Stuttgart, MSc Engineering Cybernetics study program – I

In the following we present the results of the tasks described above. We start with presenting the results of the square-wave pattern task and continue with the sawtooth pattern task and their respective cover stories.

Square wave pattern task

Regarding the first question in each set of questions, Table 3 shows that performance for the square wave pattern was poor, independent of the cover story. For analyzing the results, we first look at the average results. Then we focus on the result of the Control Group and finally we compare the results of the two Experimental Groups.

Criterion	Square Wave pattern Average Engineering Cybernetics	Square wave pattern *Bathtub* cover story (Control Group) Engineering Cybernetics	Square wave pattern *Application* cover story Engineering Cybernetics	Square wave pattern *Vehicle* cover story Engineering Cybernetics
1. When the inflow exceeds the outflow, the stock is rising.	0.82	0.88	0.71	0.71
2. When the outflow exceeds the inflow, the stocks is falling.	0.81	1.00	0.71	0.57
3. The stock should not show any discontinuous jumps (it is continuous).	0.91	0.88	0.86	0.86
4. The peaks and troughs of the stock occur when the net flow crosses zero (i.e., at t= 4,8,12,16)	0.77	0.88	0.71	0.57
5. During each segment the net flow is constant so the stock must be rising (falling) linearly.	0.81	1.00	0.71	0.57
6. The slope of the stock during each segment is the net rate (i.e., +/-25 units/time period).	0.59	0.63	0.43	0.57
7. The quantity added to (removed from) the stock during each segment is 100 units, so the stock peaks at 200 units and falls to a minimum of 100.	0.64	0.63	0.57	0.57
Mean for all items	0.76	0.84	0.67	0.63

Table 3: Performance on the square wave pattern tasks with the Bathtub, Application, and Vehicle cover stories – students of the 'BSc Engineering Cybernetics' and 'MSc Engineering Cybernetics' study program.

With 76%, the overall result was fairly similar to results of the initial study by Booth Sweeney and Sterman (2000) in which 77% of the participants answered correctly. In the following, we focus on the results of the Control Group. When comparing University of Stuttgart students' performance on the Bathtub cover story and the square wave pattern to that of MIT's students (Booth Sweeney & Sterman, 2000) we notice that performance was similar with 84% being correct (MIT: 83%). Thus, the average performance is higher than in other studies like Kapmeier, Happach and Tilebein (2014) who observed an average performance of 47% with 39 students and Ossimitz' (2002) study, who observed an average performance of 42% with 154 participants. Just like MIT students, the University of Stuttgart students did best on stating correctly that the stock does not show any discontinuous jumps (88%, compared to 96% with MIT). Considerable fewer students stated correctly the rising (88%, MIT: 87%) and falling (100%, MIT: 86%) of the stock at the appropriate times. Performance was poorer when stating correctly that the

slope of the stock during each segment is the net rate (63%, MIT: 73%), and similarly 63% (MIT: 68%) stated correctly the quantity added to or removed from the stock. It can be noted that, while average performance was poor, students' relative strengths and weaknesses on the different coding criteria is similar both to MIT results and throughout the three cover stories with criteria 2 and 5 being the best and criteria 6 and 7 the worst.

Generally, it can be stated that overall performance of students who were working on the square wave pattern with the Application cover story is higher (67%) than performance with the Vehicle story (63%). Just like with the Control Group, students of the two Experimental Groups did best on stating correctly that the stock does not show any discontinuous jumps (86% Application cover story and Vehicle cover story). The majority of the students from Experimental Group 1 also stated correctly the rising (71%) and falling (71%) of the stock at the appropriate times. This is lower than with the classic Bathtub story (stock rising: 88% and stock falling: 100%). Yet, while students with the Vehicle cover story performed similarly to participants in the Control Group indicating that the stock is rising (71%), they had difficulties in stating correctly when the stock is falling (57%). Worst performance on all three cover stories is on stating the correct slope of the stock (Bathtub: 63%; Application: 43%; Vehicle: 57%)

Interestingly, while Sterman (2002) observed pattern matching as an often occurring error in the MIT group's results, nobody of the University of Stuttgart students matched the pattern for the stock to the inflow for the square wave pattern. Also, two of the students raised concerns with the formulation of the vehicle cover story. According to their remark, the description and the figure were unclearly worded. In contemporary physics school books, tasks concerning acceleration and braking are not formulated like in our question. In physics school books especially the area of mechanics focusses on the impact of acceleration forces, traction, the direction of movements and braking forces. Thus, school's curriculum teaches Newton's second and third law and the rate-time-distance equation. What also confused the students was that in the graphical representation, the car is constantly braking while accelerating – a phenomenon that is not usually happening, except for on race tracks, maybe. As stated above (measure 3a), we had already identified the vehicle cover story as not to fit well already in our previous study (Kapmeier et al., 2014). There, we had noted the difference in tangibility and countability of the Bathtub, Application, Vehicle, and Cash flow cover stories. In addition, the misunderstanding way of representation had not appeared to us.

In addition, one student, while working on the second question understood how question one was to be answered correctly (see Figure 3). The individual states that "in Question 1 it was not clear that the braking was constant. Then the behavior would be like this:", sketching the correct behavior of the first question below the note. In addition, this also shows the rigorous time-split that we have applied in this question series.

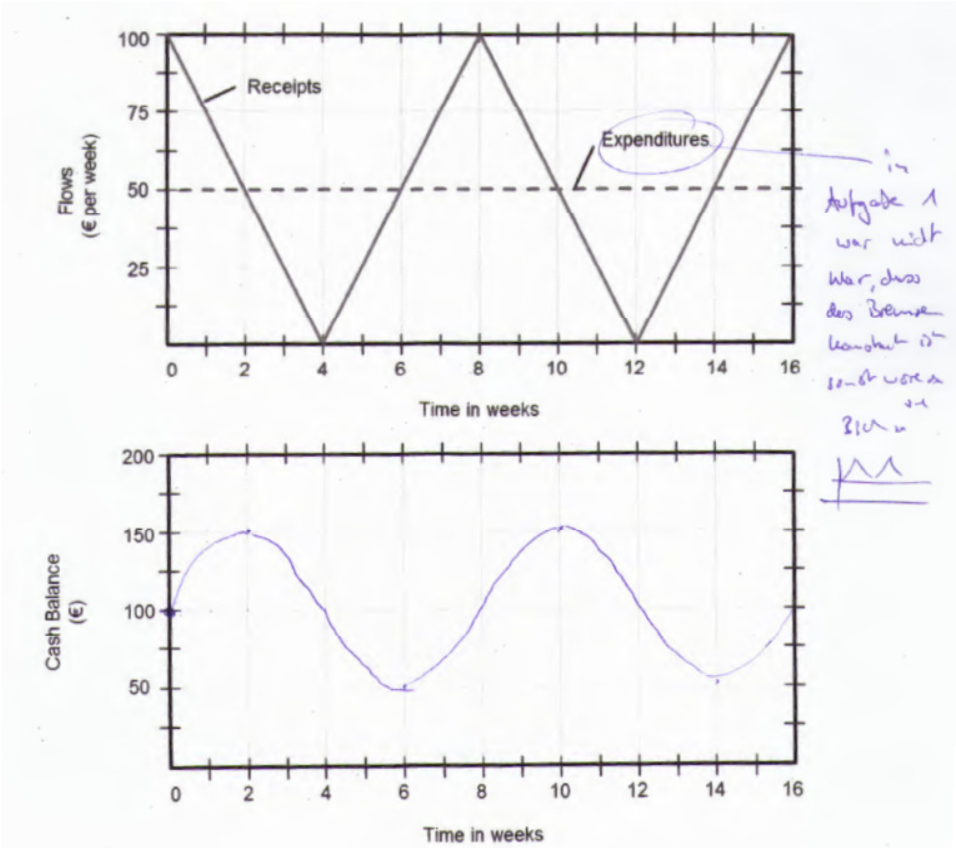


Figure 3: Notes from a participant providing the correct answer for Question 1 while working on Question 2. The student noted that "in Question 1 it was not clear that the braking was constant. Then the behavior would be like this"

These remarks had an impact on our analysis as the Vehicle cover story was especially developed for testing participants' performance on these tasks who have a knowledge domain in engineering. As the target group at the University of Stuttgart consisted of students with an engineering background, we realized that the result of this set of participants will not be fully representative. We thus only very briefly describe the performance on the sawtooth pattern in the following.

Sawtooth pattern task

Regarding the first question in each set of questions, Table 4 shows that performance for the sawtooth pattern was poor. In the following, we first look at the average results. Then we focus on the result of the Control Group and finally we compare the results of the two Experimental Groups.

Criterion	Sawtooth pattern Average Engineering Cybernetics	Sawtooth pattern Previous cover story: Bathtub (Control Group) Engineering Cybernetics	Sawtooth pattern Previous cover story: Application Engineering Cybernetics	Sawtooth pattern Previous cover story: Vehicle Engineering Cybernetics
1. When the inflow exceeds the outflow, the stock is rising.	0.82	0.75	0.86	0.86
2. When the outflow exceeds the inflow, the stock is falling.	0.77	0.75	0.71	0.86
3. The stock should not show any discontinuous jumps (it is piecewise continuous).	1.00	1.00	1.00	1.00
4. The peaks and troughs of the stock occur when the net flow crosses zero (i.e., t = 2, 6, 10, 14).	0.77	0.75	0.71	0.86
5. The slope of the stock at any time is the net rate. Therefore: a. when the net flow is positive and falling, the stock is rising at a diminishing rate (0<t<2; 8< t<10). b. when the net flow is negative and falling, the stock is falling at an increasing rate (2<t<4; 10<t<12). c. when the net flow is negative and rising, the stock is falling at a decreasing rate (4<t<6; 12<t<14). d. when the net flow is positive and rising, the stock is rising at an increasing rate (6<t<8; 14<t<16).	0.55	0.50	0.57	0.57
6. The slope of the stock when the net rate is at its maximum is 50 units/period (t = 0, 8, 16).	0.46	0.38	0.57	0.43
7. The slope of the stock when the net rate is at its minimum is -50 units/period (t = 4, 12).	0.46	0.38	0.57	0.43
8. The quantity added to (removed from) the stock during each segment of 2 periods is the area of the triangle bounded by the net rate, or +/-(1/2)*50 units/period*2 periods = 50 units. The stock therefore peaks at 150 units and reaches a minimum of 50 units.	0.56	0.25	0.86	0.57
Mean for all items	0.67	0.59	0.73	0.70

Table 4: Performance on the sawtooth pattern tasks, differentiated according to the previous Bathtub, Application, and Harvester cover stories – students of the 'BSc Engineering Cybernetics' and 'MSc Engineering Cybernetics' study program.

With 67%, the overall result was lower than results of the initial study by Booth Sweeney and Sterman (2000) in which 77% of the participants answered correctly. About five-sixth of the University of Stuttgart students showed correctly that the stock rises when the inflow exceeds the outflow (and three-quarter of them correctly state that the stock falls when the outflow exceeds the inflow). 75% (75%) answered correctly when the previous cover story was the Bathtub story, 86% (71%) the Application story, and 86% (86%) the Vehicle story. Further, more or less the same share of participants from the three groups also marked the peaks and troughs of the stock at the appropriate points in time (75% of Bathtub group, 71% of the Application group, and 86% of the Vehicle group). A fairly high number of University of Stuttgart students (45%) failed to relate the net rate to the stock (50% Bathtub, 43% Application, and 43% Vehicle). Whereas about half of MIT students correctly drew the maximum (52%) and the minimum (51%) slope of the stock, the average was lower in the University of Stuttgart study (46% maximum and 46% minimum). However, when looking at the three groups, we observe great differences. Only 38% (maximum slope) and 25% (minimum slope) of the students who had worked on the Bathtub story answered this correctly. This is much less than the students who had previously worked on the Application story (57% and 86%) and the Vehicle story (43% and 57%). Everybody recognized that there are no discontinuous jumps in the stock in the 'Cash flow' task (100% for all participants, regardless of the previous cover story), which very much equals MIT students' performance (99%). Whereas more than half (56%) of the University of Stuttgart students correctly calculated the maximum and minimum of the stock, we again note differences in the performance of the individual previous cover stories. Those who had worked previously on the Application task did best (73%), followed by those who had worked on the

Vehicle task (70%), and finally, only 59% of those who had worked on the Bathtub story, stated this correctly. To do so correctly, students simply had to calculate the area of the triangle bound by the net rate.

As a consequence of the feedback received by the students, we developed the new cover story for the engineering knowledge domain that describes the harvest of wheat, as shown in Figure 4. It describes how many kilograms of wheat a harvester harvests and how much it ejects over a time horizon of 16 minutes. There is an initial quantity of 100kg of wheat in the harvester that is processed. Along with this cover story goes the square-wave pattern as shown in Booth Sweeney and Sterman (2000).

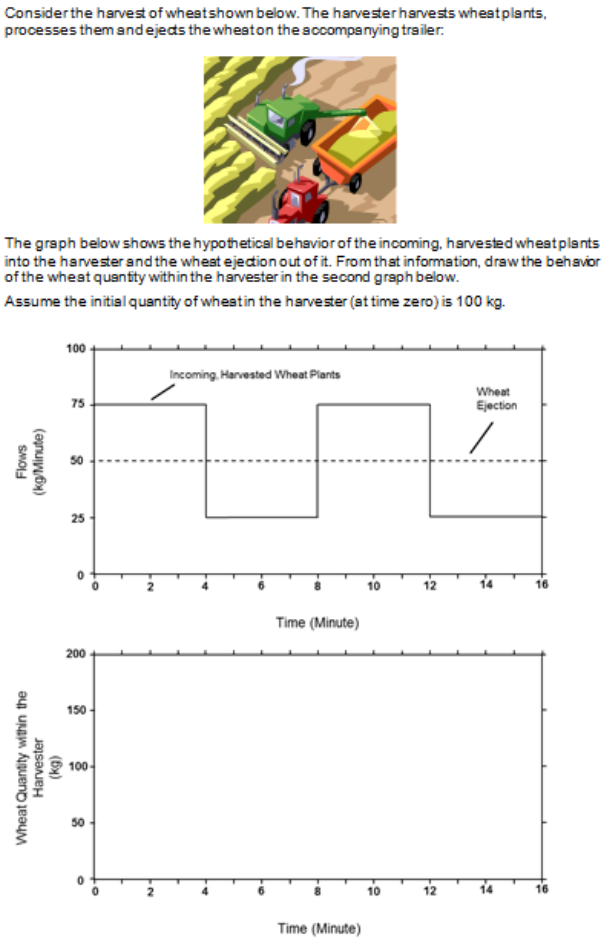


Figure 4: The new cover story relating to the engineering knowledge domain describes the process in a harvester

The new cover story also fits better into the observations on countability and tangability in the other cover stories (Kapmeier et al., 2014), as can be seen in the Table 5. Like the Bathtub cover story, the Harvester cover story deals with tangible stuff, describing water flowing into and out of a bathtub and wheat flowing into and out of a harvester. In addition, wheat corns can be counted, which goes along with

money that can be counted in the Cash flow story. The Harvester cover story thus has more overlap in tangibility and countability with the original cover stories than the Vehicle cover story.

Cover story	Tangibility	Countability
Bathtub	●	○
Application	◒	●
Vehicle	n.a.	n.a.
Harvester	●	◒
Cash flow	◒	●

Table 5: Flows differ in terms of tangibility and countability for the different cover stories

We tested the new question set in January 2015 with sixteen Engineering students on Bachelor, Master, and PhD level of the University of Stuttgart. When asked about their understanding of the Harvester cover story the participants fed back that they understood the question well. There was no hint that the question is difficult to understand. This was a confirmation to us the inflow into the harvester, outflow out of the harvester and the stock of wheat in the harvester make sense as the cover story to the square wave pattern in the engineering knowledge domain.

The new, tested set of questions was then handed out in an introductory class to cybernetics in the first semester of the BSc Engineering Cybernetics study program. Fifty-eight students participated in this session. As can be seen from the Table 6, the background data sheet asked for the same information as before, including the newly added information.

The proportion of male and female participants is unevenly distributed in all groups with a majority of male students (86%). As can be seen from the Table 6, all subjects were in their first semester. All but one participants were younger than 24. The particular student who was between 25 and 30 years old has also finished a previous program in the engineering field – thus in the same knowledge domain. Only a small number of participants had done vocational training (5%) or internships (7%). Moreover, the knowledge domain was in the same (engineering) or related (IT) areas. A few participants originate from outside Europe (5%). One participant originates from Asia and the Middle East and two from Africa. 95% of the participants originate from Europe. German, the study program language, was the first language for the majority of participants (91%). None of the participants had played the beer distribution game before.

The high number of participants with an engineering background met our objective to study the effect of previous knowledge domain with comparing it to our previous study in which we studied 39 students (Kapmeier et al., 2014).

	Uni Stuttgart Winter 2014/2015 II Engineering Cybernetics			
Task	**Total numbers**	**Control Group: Bathtub and Cash Flow**	**Experimental Group 1: Application and Cash Flow**	**Experimental Group 2: Harvester and Cash Flow**
Total Number of Students	**58**	**19**	**18**	**21**
Age				
19-24	57	18	18	21
25-30	1	1	0	0
31-35	0	0	0	0
36 and up	0	0	0	0
Gender				
Male	50	16	14	20
Female	8	3	4	1
Student Status				
Technical Cybernetics	58	19	18	21
TeMa	0	0	0	0
Prior Field of Study				
Business/Management	0	0	0	0
Engineering	1	0	0	1
Social Science	0	0	0	0
Computer Science	0	0	0	0
Mathematics	0	0	0	0
Humanities	0	0	0	0
Highest Prior Degree				
BA	0	0	0	0
BSc	0	0	0	0
MA, MSc, Diplom	0	0	0	0
Ph.D.	0	0	0	0
High School	57	19	18	20
BE. JD, BBA, MD, CPA	1	0	0	1
Current Field of Study				
Business/Management	0	0	0	0
Engineering	58	19	18	21
Social Science	0	0	0	0
Science	0	0	0	0
Computer Science	0	0	0	0
Mathematics	0	0	0	0
Humanities	0	0	0	0
Vocational training				
Business/Management	0	0	0	0
Engineering	3	1	0	2
IT	0	0	0	0
Other	0	0	0	0
No vocational training	37		18	19
Internship (=> 3 months)				
Business/Management	0	0	0	0
Engineering	2	1	0	1
IT	2	0	0	2
Other	3	3	0	0
No vocational training	51	15	18	18
Region of Origin				
North America (Aus. + NZ)	0	0	0	0
Europe	55	18	17	20
Asia and Middle East	1	1	0	0
Latin and South America	0	0	0	0
Africa	2	0	1	1
First language				
German	53	17	16	20
English	0	0	0	0
Other	5	2	2	1
Teaching Language				
English	6	2	1	3
German	51	16	17	18
Beer Game Experience				
Played before	0	0	0	0
Have not played	58	19	18	21

Table 6: Subject demographics – University of Stuttgart, BSc Engineering Cybernetics study program II

In the following we present the results of the first semester Engineering Cybernetics students of the University of Stuttgart. As above, we start with presenting the results of the square-wave pattern task and continue with the sawtooth pattern task and their respective cover stories.

Square wave pattern task

Regarding the first question in each set of questions, Table 3 shows that performance for the square wave pattern was relatively poor, independent of the cover story. For analyzing the results, we first look at the average results. Then we focus on the result of the Control Group and finally we compare the results of the two Experimental Groups.

With 83%, the overall result was slightly higher than results of the initial study by Booth Sweeney and Sterman (2000) in which 77% of the participants answered correctly.

In the following, we focus on the results of the Control Group in order to find support for or against the hypothesis stated above. When comparing University of Stuttgart students' performance on the Bathtub cover story and the square wave pattern to that of MIT's students (Booth Sweeney & Sterman, 2000) we notice that performance was similar with 86% being correct (MIT: 83%). Thus, the average performance is higher than in the study carried out with higher semester students of the same study program, as described above. Just like MIT students, the University of Stuttgart students did best on stating correctly that the stock does not show any discontinuous jumps (89%, compared to 88% with Stuttgart I and 96% with MIT). Similarly well, students stated correctly the rising (Stuttgart II: 89%, Stuttgart I: 88%, MIT: 87%) and falling (Stuttgart II: 89%, Stuttgart I: 100%, MIT: 86%) of the stock at the appropriate times. Performance was higher when stating correctly that the slope of the stock during each segment is the net rate (Stuttgart II: 84%, Stuttgart I: 63%, MIT: 73%), and similarly (Stuttgart II: 79%, Stuttgart I: 63%, MIT: 68%) stated correctly the quantity added to or removed from the stock. It can be noted that, while average performance was relatively poor, students' relative strengths and weaknesses on the different coding criteria is similar to both Stuttgart I and MIT results and throughout the three cover stories with criteria 2 being the best and criteria 6 and 7 the worst.

Interestingly, unlike above where it was clearly visible where the best and worst coding criteria were for the different groups, here performance on the different coding criteria is very close together, with a maximum difference of 10 percent points. In the Harvester cover story, for example, the worst criteria is the slope of the stock (76%) and the best criteria are 1 (raising stock), 2 (falling stock), and 4 (peaks and troughs) with 86%. Generally, it can be stated that overall performance of students who were working on the square wave pattern with the Harvester cover story is slightly higher (82%, Vehicle story with Stuttgart I: 63%) than performance with the Application story (80%; Application cover story with Stuttgart I: 67%).

Results differ between the cover stories on which criterion the participants did best. Students working on the Application cover story did best on stating that the stock is rising (falling) when the inflow (outflow) exceeds the outflow (inflow) with 83% (83%) and that the stock does not show any discontinuous jumps (also 83%). Students working on the Harvester cover story did best also on stating that the stock is rising (falling) when the inflow (outflow) exceeds the outflow (inflow) with 86% (86%). Interestingly, students perform well on criterion 4, stating that the peaks and troughs occur when the net flow crosses zero (86%). The percentage is similar to the result of the Control group (84%) and higher than with the Application story (78%).

Criterion	Square Wave pattern Average Engineering Cybernetics	Square wave pattern *Bathtub* cover story (Control Group) Engineering Cybernetics
1. When the inflow exceeds the outflow, the stock is rising.	0.84	0.84
2. When the outflow exceeds the inflow, the stocks is falling.	0.86	0.89
3. The stock should not show any discontinuous jumps (it is continuous).	0.85	0.89
4. The peaks and troughs of the stock occur when the net flow crosses zero (i.e., at t= 4,8,12,16)	0.83	0.84
5. During each segment the net flow is constant so the stock must be rising (falling) linearly.	0.83	0.89
6. The slope of the stock during each segment is the net rate (i.e., +/-25 units/time period).	0.79	0.84
7. The quantity added to (removed from) the stock during each segment is 100 units, so the stock peaks at 200 units and falls to a minimum of 100.	0.79	0.79
Mean for all items	0.83	0.86

		correlation with Control Group				correlation with Control Group		
	Square wave pattern *Application* cover story Engineering Cybernetics	X^2	*Cramer's V*	*p*	Square wave pattern *Harvester* cover story Engineering Cybernetics	X^2	*Cramer's V*	*p*
1.	0.83	0.720	0.200	0.396	0.86	0.825	0.208	0.364
2.	0.83	1.800	0.316	0.180	0.86	0.419	0.149	0.517
3.	0.83	1.800	0.316	0.180	0.81	0.596	0.177	0.440
4.	0.78	0.257	0.120	0.612	0.86	0.825	0.208	0.364
5.	0.78	1.004	0.236	0.316	0.81	0.596	0.177	0.440
6.	0.78	0.257	0.120	0.612	0.76	0.090	0.069	0.764
7.	0.78	0.023	0.036	0.880	0.81	0.048	0.050	0.827
	0.80				0.82			

Table 7: Performance on the square wave pattern tasks with the Bathtub, Application, and Harvester cover stories – students of the 'BSc Engineering Cybernetics' study program. The X^2 statistic tests the hypothesis that performance on the two treatment conditions is the same

For the Harvester cover story, students performed worst on identifying that the slope of the stock is the net rate during each segment (76%), compared to 78% (Application) and 84% (Bathtub). For the Application cover story, students performed worst in criteria four until seven with 82% each. Yet, this is still higher than with the Stuttgart I group, with performance ranging between 43% (criterion 6) and 71% (criteria 4 and 5).

In contrast to other studies (i.e.,(Kapmeier et al., 2014)), pattern matching was not a typical approach to solution. Only one participant working on the Bathtub and the Harvester story, and two individuals in the Application story matched the pattern for the stock to the inflow. Because of the relatively high solution percentage of all criteria, all criteria correlate highly and significantly with each other, as observed before (Kapmeier, 2004; Kapmeier et al., 2014). It can, for example, be stated here that the first criteria correlate highly (Bathtub: Pearson's R=0.792, Application: Pearson's R=1, and Harvester: Pearson's R=1) and significantly (Bathtub: p=0.000, Application: p=0.000, and Vehicle: p=0.000) with each other.

Sawtooth pattern task

As in the previous surveys the three study groups on average found the task with the sawtooth pattern for the inflow and the underlying Cash flow cover story more difficult than the square wave pattern. As can be seen from the Table 8, average performance of the University of Stuttgart students on the Cash flow task was poor (60%). Interestingly, there are somewhat large deviations from average when looking at performances between the cover stories: students who had previously worked on the Bathtub cover story perform worst (Control Group: 48%), followed by Experimental Group 1 with the Application cover story (64%), and Experimental Group 2 with the Harvester cover story (69%). When compared to previous studies, in which 45% (Kapmeier, 2004), 51% (Booth Sweeney & Sterman, 2000), and 48% (Ossimitz, 2002) correctly answered the Cash flow task with the underlying sawtooth pattern, performance is in the same range.

	Criterion	Sawtooth pattern Average Engineering Cybernetics	Sawtooth pattern Previous cover story: Bathtub (Control Group) Engineering Cybernetics
1.	When the inflow exceeds the outflow, the stock is rising.	0.62	0.53
2.	When the outflow exceeds the inflow, the stock is falling.	0.63	0.58
3.	The stock should not show any discontinuous jumps (it is piecewise continuous).	0.85	0.79
4.	The peaks and troughs of the stock occur when the net flow crosses zero (i.e., t = 2, 6, 10, 14).	0.63	0.58
5.	The slope of the stock at any time is the net rate. Therefore: a. when the net flow is positive and falling, the stock is rising at a diminishing rate (0<t<2; 8< t<10). b. when the net flow is negative and falling, the stock is falling at an increasing rate (2<t<4; 10<t<12). c. when the net flow is negative and rising, the stock is falling at a decreasing rate (4<t<6; 12<t<14). d. when the net flow is positive and rising, the stock is rising at an increasing rate (6<t<8; 14<t<16).	0.57	0.42
6.	The slope of the stock when the net rate is at its maximum is 50 units/period (t = 0, 8, 16).	0.52	0.32
7.	The slope of the stock when the net rate is at its minimum is -50 units/period (t = 4, 12).	0.52	0.32
8.	The quantity added to (removed from) the stock during each segment of 2 periods is the area of the triangle bounded by the net rate, or +/-(1/2)*50 units/period*2 periods = 50 units. The stock therefore peaks at 150 units and reaches a minimum of 50 units.	0.50	0.32
	Mean for all items	0.60	0.48

		with Bathtub cover story correlation with Control Group				with Bathtub cover story correlation with Control Group		
	Sawtooth pattern Previous cover story: Application Engineering Cybernetics	**X^2**	***Cramer's V***	***p***	**Sawtooth pattern Previous cover story: Harvester Engineering Cybernetics**	**X^2**	***Cramer's V***	***p***
1.	0.61	3.378	0.433	0.066	0.71	0.148	0.088	0.701
2.	0.61	1.606	0.299	0.205	0.71	0.012	0.025	0.912
3.	0.89	1.800	0.316	0.180	0.86	1.127	0.244	0.288
4.	0.61	1.606	0.299	0.205	0.71	0.012	0.025	0.912
5.	0.61	1.169	0.255	0.280	0.67	0.224	0.109	0.636
6.	0.61	1.870	0.322	0.171	0.62	0.652	0.185	0.419
7.	0.61	1.870	0.322	0.171	0.62	0.652	0.185	0.419
8.	0.56	0.112	0.079	0.737	0.62	0.046	0.049	0.829
	0.64				0.69			

Table 8: Performance on the sawtooth pattern tasks, differentiated according to the previous Bathtub, Application, and Harvester cover stories – first semester students of the 'BSc Engineering Cybernetic' study program. The X^2 statistic tests the hypothesis that performance on the two treatment conditions is the same

Only slightly more than half of the University of Stuttgart students showed correctly that the stock rises (falls) when the inflow exceeds the outflow (and vice versa): 53% (58%) answered correctly when the previous cover story was the Bathtub story, 61% (61%) the Application story, and 71% (71%) the

Harvester story. Further, more or less the same share of participants from the three groups who succeeded in the criteria also marked the peaks and troughs of the stock at the appropriate points in time (58% of Bathtub group, 61% of the Application group, and 71% of the Harvester group). A fairly high number of University of Stuttgart students (43%) failed to relate the net rate to the stock (58% Bathtub, 39% Application, and 23% Harvester). Whereas about half of MIT students correctly drew the maximum (52%) and the minimum (51%) slope of the stock, the average was almost equal in the University of Stuttgart study (52% maximum and minimum as well). However, when looking at the three groups, we observe great differences. Only 32% of the students who had worked on the Bathtub story answered this correctly. This is only about half of the students who had previously worked on the Application story (61% for both) and the Harvester story (62%). Whereas a large share of students who had previously worked on the Harvester story (86%) and the Application story (89%) correctly recognized that there are no discontinuous jumps in the stock in the 'Cash flow' task, only 79% of those who had worked on the Bathtub story did this correctly. The numbers of the Application and Harvester students do not reach MIT students' performance (99%). Whereas more than half (56%) of the University of Stuttgart students who had worked previously on the Application task and 62% of those having worked on the Harvester story correctly calculated the maximum and minimum of the stock, only 32% did so in the Bathtub story. To do so, students simply had to calculate the area of the triangle bound by the net rate.

As in the square wave task, pattern matching was a more common error for the groups (21% in the Bathtub group, 27% in the Application group, and 19% in the Harvester group). While this is higher than in the square wave pattern task, these numbers are small compared to our previous study (Kapmeier et al., 2014) and the MIT study (Booth Sweeney & Sterman, 2000). Also, as in the square wave task, many criteria correlate highly and significantly with each other in the sawtooth task, the strongest correlation being between criteria 1 (rising stock) and 2 (falling stock) (previous cover stories: Bathtub: Pearson's R=0.899 and p=0.000; Application and Harvester: Pearson's R=1), and between criteria 1, 2 and 4 (peaks and troughs) (Bathtub: Pearson's R=0.899 with p=000 for 1 and 4 and R=1.000 for 2 and 4 Pearson's R=1.000; Application and Harvester: Pearson's R=1.000 for 1 and 2 with 4).

Hence, one might assume that the subjects who follow the rule of an increasing stock when the inflow exceeds the outflow in the square wave task will do the same in the sawtooth task, independent from the cover story. So, there should be a correlation between criterion 1 in the Bathtub square wave and the sawtooth tasks. In fact, the correlation between criterion 1 in the square wave and sawtooth task is significant (Bathtub cover story: Pearson's R=0.456; Application cover story: Pearson's R=0.561). Data for the Harvester cover story reveal that there is no correlation: Pearson's R=0.344). Likewise, the relation between criteria 2 in the respective tasks for the Bathtub square wave and Cash flow sawtooth (Pearson's R= 0.344) and the Harvester square wave and the Cash flow sawtooth are not correlating (Pearson's R=

0.402). However, criteria 2 are correlating for the Application square wave and Cash flow sawtooth (Pearson's R=0.561).

Discussion of Results

In our previous study we argued that performance on SF-tasks increases when the problem context is embedded in the problem solver's knowledge domain. With this study we explore the propositions with participants in the engineering domain. Although we again find our hypothesis supported to some extent, there are still aspects that have to be further explored.

We have conducted two surveys with engineering students, one in November 2014 with higher semester students and a second in January 2015 with first semester students. In the following we first discuss results of the new data. Then we put them into relation to our previous data from business students. Note that because of ambiguous formulation in the description of the Vehicle cover story, we exclude analysis for the higher semester engineering students from this comparison with our previous results.

Overall, it can be noted that first semester engineering cybernetics students (January 2015) performed considerably better than the higher semester students (November 2014) in the first task. Average performance on the square wave pattern task was 76% with the higher semester students, compared to 83% with the first semester cybernetics students. When looking at the three different cover stories, again the first semester students perform better (Bathtub: 86%; Application: 80%; Harvester: 82%) than higher semester students (Bathtub: 84%; Application: 67%; Vehicle: 63%) with each cover story. This is an interesting observation and we should test this further. One could argue that the Vehicle cover story was ambiguously formulated and thus the results are not comparable.

However, the two tasks are independent from each other, which allows us to look at the more difficult sawtooth pattern task with the Cash flow story. Here, overall performance of higher semester students (67%) is higher than with first semester students (60%). In terms of performance gap between first and higher semester students per cover story, we observed differences: With the Bathtub as previous cover story, higher semester students perform considerably better in the sawtooth task (59% vs. 48%), the gap is less with the previous Application cover story (73% vs. 64%) and, interestingly, quite narrow with the previous Vehicle/Harvester cover story (70% vs. 69%).

An initial explanation for this observation could be that the square wave task asks for simple graphical integration – a topic that first semester students are still more familiar with than higher semester students: Graphical integration is part of A-level examination in Germany and first semester students have just passed their A-level exams. Moreover, higher semester students are more used to task presentation that conform to their study programs' or disciplines' typical means of representation and visualization.

Maybe they try to use the tools they had learned to solve the Bathtub Dynamics tasks, which does not lead to the wished-for results. In other words, with advancing studies, students may be increasingly mentally focused on their acquired problem solution methods. An interdisciplinary jump seems to be out-of-range. At the same time, in the beginning of their studies, students might seem more open and free of disciplinary mindsets.

Moreover, pattern matching was a problem-resolution mode for only a few participants. Reason for this might be that BSc students in engineering cybernetics are highly mathematically focused. This observation needs to be confirmed with future studies.

As the January 2015 study with first semester students used the new engineering-specific Harvester cover story, we refer to these data in the following analysis. The Table 9 summarizes the respective expectations and findings, where potential bias indicates the need for further research.

		confirmed	supported	not supported	contradicted	biased?
The Experiment	hypothesis at large		x			x
	comparable cover stories		x			
	alignment with original study		x			
	alignment with other studies		x			
Results of the Engineering Students	Overall performance in Harvester cover story better than in Application cover story		x			x
	Overall performance in Harvester cover story better than in original Bathtub cover story				x	x
	Overall performance in Bathtub cover story better than in Application cover story		x			
	Relatively better performance in complicated task that is engineering related than in easy task from distant knowledge domain	not applicable				
	SF-performance (avoiding pattern matching) is much better in Harvester vs. Bathtub vs Application cover stories			x		x
	Sawtooth task is more difficult, just like in previous studies, leading to lower average performance than in square wave task		x			
	All three groups, as they have the same background, perform equally well on sawtooth task (second task), independent from first task			x		
	Pattern matching is most common error in sawtooth task, just like in previous studies				x	x

Table 9: Expectations and findings of the present study with engineering students

In addition, one of our findings regarding methodology was that the precise time split has worked. This way we ensure that all participants spend the same time for each question.

In the following we compare the results of this study to the previous study conducted with students who have a business background (see Table 1). First, it can be stated that the overall performance of engineering students is much higher than performance of students with a knowledge domain in business (Kapmeier et al., 2014). Note that because the two different tasks within one survey are independent from each other, the square wave pattern task can be used for our analysis. Average performance with business students in the square wave pattern task was 49%. With our engineering groups, performance on the square wave pattern task was 76% and 83% - this is 55% and nearly 70% better than the business students. The same holds for the sawtooth pattern task. Here, performance of the business students was 38%; first semester engineering students performed on average 60% correctly, thus more than 50% better than business students. Even in the questions that refer to a business knowledge domain, the engineering students performed better than the business students. An explanation could be that the BSc in Engineering Cybernetics are generally very good students, better than their peers also in the engineering faculty. For example, they write some of their exams together with engineering students of other engineering fields and they usually perform above-average. This insight might relativize our findings which we will have to support in future studies.

Furthermore, there are limitations to our study with respect to the comparability of the previous study with business students (Kapmeier et al., 2014) and the study presented here which was conducted with engineering students. Our previous study was conducted before the new engineering cover story (Harvester) and the new procedure were introduced. This weakens comparability. Moreover, not only participants' educational background differs in terms of their study program. Their demographics varies between the two studies in terms of percentage of native speakers of the task language, share of female students, variety of nationalities, and the level of domain-specific experience.

General Discussion and Further Research

In our study we argue that domain specific experience and familiarity with the problem context play a role in SF-performance. To test this hypothesis we have varied the cover stories that go along with the square wave pattern of the classic Bathtub dynamics study (Booth Sweeney & Sterman, 2000). We have found support for this hypothesis in our previous study with business students (Kapmeier et al., 2014). In this paper we test our hypothesis with participants of an engineering background. We collected data of 22 participants in the first round (November 2014) and 58 participants in the second round (January 2015). Also these data support our hypothesis. Yet, although the results are promising, there are some limitations and a number of issues remain to be further explored, e.g. with respect to the studies' data base, and the

phenomenon of SF-failure as well as a general discussion on skill-building in dealing with SF-structures, respectively.

Two of these issues are related to the data we have collected so far: Firstly, with respect to range of participants, we intended to analyze the effect of previous experience from vocational training or internships collected in a different field than the field of study on performance. As an example, a participant who had done a business-related vocational training is studying engineering. Our suggestion is that this individual would understand the business-related task better than somebody who is studying engineering without having done this vocational training in a different knowledge domain. Although we included a number of related additional questions into the background data sheet, our study did not reveal a significant number of participants with relevant domain-specific experience from other sources and in other domains. This is why this path of research has to be further explored, e.g. by conducting studies with participants from further education studies at university level who come with a background in e.g. vocational training, or with tenure in a specific firm or position. In addition, it could be interesting to analyze whether differences in SF-performance could be related to the source of the respective domain-specific knowledge or the mode of acquisition, for example, from hands-on experience vs. from a study program's rather theoretical perspective.

Secondly, our first study's (Kapmeier et al., 2014) limitation in number of participants still holds true, as we focused on engineering students in our new study, while we conducted the first study with business students only. Hence in the next steps the number of participants in the business knowledge domain should be increased.

Other issues relate to the phenomenon of SF-performance as such: In the following we focus on three aspects with respect to SF-performance, its assessment, and acquisition of related skills. Firstly, we find support for our hypothesis from two comparable sets of participants from different knowledge domains. Yet, the average performance is 83% on the square wave pattern and 60% on the sawtooth pattern. Average performance on the square wave pattern is (slightly) higher than the best results from previous research. This raises the question, especially for the square wave pattern, whether performance is really 'poor'. This finding could contribute to a discussion about the aspiration level in SF-performance.

Secondly, as motivation can act as a driver of problem-solving performance in general, one could think of different ways of motivation. The original study and subsequent ones underlined that students were not graded nor paid or otherwise rewarded for their test scores or participation. However, while this voluntary character may hold true for everyday situations, dealing with SF-structures in professional contexts may involve more extrinsic aspects of motivation like, e.g. the chance of winning bonuses and other incentives, or the risk of losing a contract or even the job, on the contrary. If we take our results as a first hint towards teaching stock-and-flow systems also in domain-specific approaches and linking it more closely to professional contexts, we should also discuss a more realistic extrinsic motivation.

Moreover, as hints for further research, we suggest to further discuss whether the cash flow cover story really is general knowledge with "everyday context" (Booth Sweeney & Sterman, 2000: 252) or rather business-specific, as it relates to a company's cash balance. Cronin and Gonzalez (2007) for instance, propose to take a private bank account as an everyday task context. We therefore suggest to expand our investigation to the second (sawtooth) task by formulating other alternative cover stories from more general knowledge on the one hand and from the engineering domain on the other hand, in order to completely mirror our first study.

To sum up, our hypothesis was that SF-performance increases when the problem context is embedded in the educational background of the problem solver. This hypothesis was confirmed with participants from a business study program in our first study (Kapmeier et al., 2014). With our new study we gather additional data from the engineering domain that also supports our hypothesis. With the next steps sketched above we want to contribute to a further reaching general discussion among scholars about how to assess and improve SF-performance.

Bibliography

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

Brunstein, A., Gonzalez, C., & Kanter, S. (2010). Effects of domain experience in the stock–flow failure. *System Dynamics Review, 26*(4), 347–354.

Cronin, M. A., & Gonzalez, C. (2007). Understanding the building blocks of dynamic systems. *System Dynamics Review, 23*(1), 1–17. doi: 10.1002/sdr.356

Cronin, M. A., Gonzalez, C., & Sterman, J. D. (2009). Why don't well-educated adults understand accumulation? A challenge to researchers, educators, and citizens. *Organizational Behavior and Human Decision Processes, 108*(1), 116–130.

Fischer, H., & Gonzales, C. (in press). Making sense of dynamic systems: How our understanding of stocks and flows depends on a global perspective. *Cognitive Science*.

Größler, A., & Strohhecker, J. (2012). Tangible stock/flow experiments - Adressing issues of naturalistic decision making. *Proceedings of the 30th International Confernce of the System Dynamics Society in St. Gallen Switzerland.*

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 in Palermo, Italy.*

Kapmeier, F. (2004). Findings from four years of bathtub dynamics at higher management education institutions in Stuttgart. *Proceedings of the 22nd International Confernce of the System Dynamics Society*.

Kapmeier, F., Happach, R. M., & Tilebein, M. (2014). *Bathtub Dynamics Revisited: Does Educational Backgroud Matter?* Paper presented at the System Dynamics Conference, Delft, NL.

Kapmeier, F., & Zahn, E. O. K. (2001). Bathtub Dynamics: Results of a Systems Thinking Inventory at the Universität Stuttgart, Germany. Retrieved from www.esb-business-school.de/business-school/organisation/professoren-und-dozenten/kapmeier.html website:

Korzilius, H., Raaijmakers, S., Rouwette, E., & Vennix, J. (t.b.a.). Thinking Aloud While Solving a Stock-Flow Task: Surfacing the Correlation Heuristic and Other Reasoning Patterns. *Systems Research and Behavioral Science, t.b.a*(t.b.a.), t.b.a.

Moxnes, E., & Jensen, L. (2009). Drunker than intended: Misperception and information treatments. *Drug and Alcohol Dependence, 105*(1-2), 63–70.

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

Schwarz, M. A., Epperlein, S., Brockhaus, F., & Sedlmeier, P. (2013). Effects of illustrations, specific contexts, and instructions: Further attempts to improve stock-flow task performance. *Proceedings of the 31st International Confernce of the System Dynamics Society in Cambridge, MA, USA*.

Sedlmeier, P., Brockhaus, F., & Schwarz, M. (2014). Visual integration with stock-flow models: How far can intuition carry us? In T. Wassong, D. Frischemeier, P. R. Fischer, R. Hochmuth, & P. Bender (Eds.), *Mit Werkzeugen Mathematik und Stochastik lernen – Using Tools for Learning Mathematics and Statistics* (pp. 57–70). Wiesbaden: Springer Fachmedien.

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

Sterman, J. D. (2010). Does formal system dynamics training improve people's understanding of accumulation? *System Dynamics Review, 26*(4), 316–334.

Sterman, J. D., & Booth Sweeney, L. (2002). Cloudy skies: assessing public understanding of global warming. *System Dynamics Review, 18*(2), 207–240. doi: 10.1002/sdr.242

Appendix to:

Bathtub Dynamics Revisited:
Does Educational Background Matter?

Note:

The three "Bathtub Dynamics" sets of questions consist of the data sheet

1) Task I(a) and Task II
2) Task I(b) and Task II
3) Task I(c) and Task II

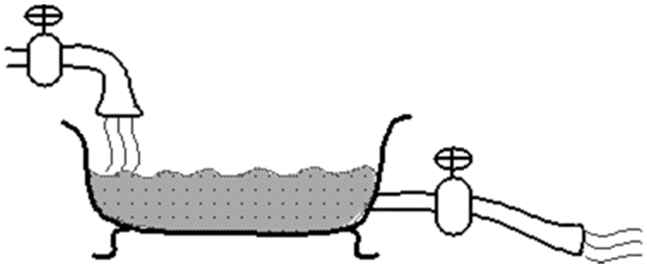
Data sheet

1. Age: _____ Gender: □ M □ F
2. School and program:
   - □ ESB IB
   - □ ESB Master IBD
   - □ ESB Master IAT
   - □ Other ESB Program _____
   - □ ESB international exchange student
   - □ University of Stuttgart TKyb
   - □ University of Stuttgart TeMa
   - □ Other (specify course and university):
3. Years of studies in general: 1 2 3 4 5
4. Semester of current program: 1 2 3 4 5 6 7 8 9
5. Have you done vocational training? □ Yes □ No
   If yes, specify profession: _____
6. Have you gained prior work-experience in same position over more than 3 months?
   □ Yes □ No
   If yes, specify work-experience: _____
7. Mother tongue is □ English □ German □ Other: _____
8. Teaching language is □ English □ German □ Other: _____
9. Country of Origin: _____
10. Highest previous degree: □ High School □ B.A. □ B.Sc. □ B.Eng. □ B.Ed. □ LL.B. □ Magister □ M.A. □ M.Sc. □ MBA □ B.Eng. □ Ph.D.
    □ Diploma □ Other _____
    Field or Major: _____
11. Have you played the 'Beer Distribution Game'? □ Yes □ No

There are two questions. You have 5 minutes to solve each.

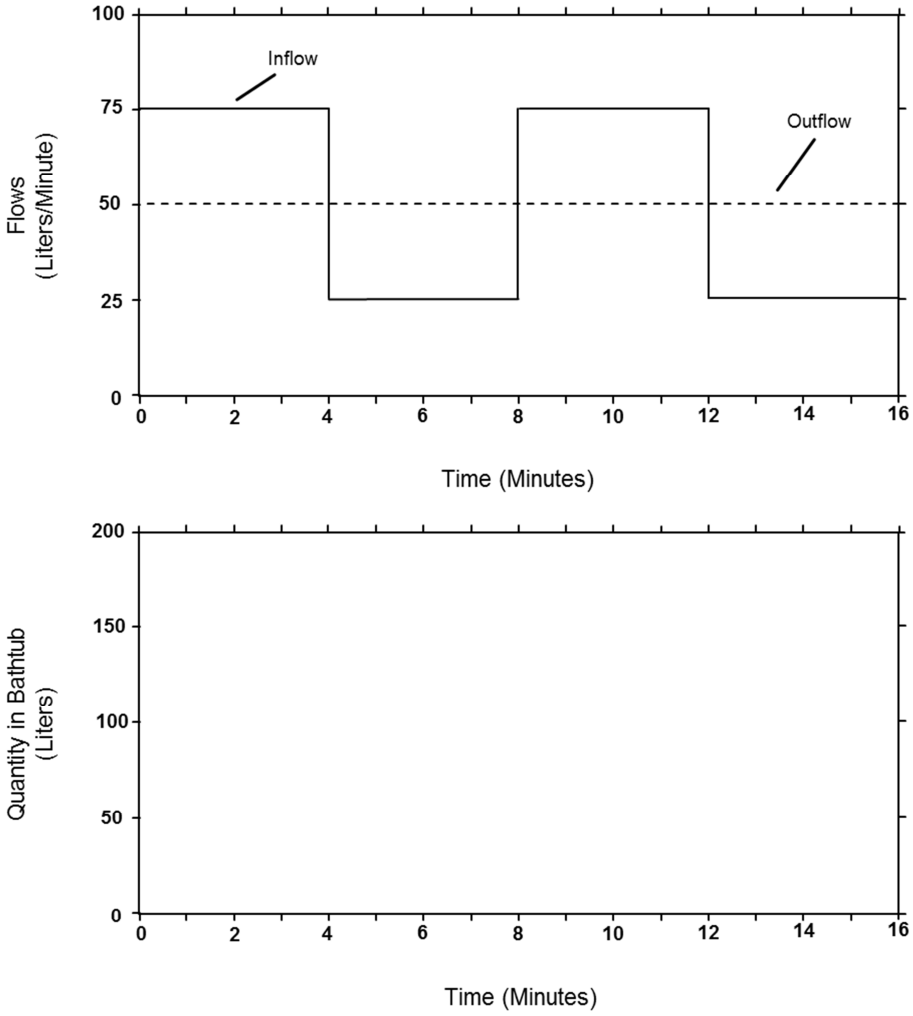
"Bathtub Dynamics" Task I (a)

Consider the bathtub shown below. Water flows in at a certain rate (on the left), and exits through the drain at another rate (on the right):



The graph below shows the hypothetical behavior of the inflow and outflow rates for the bathtub. From that information, draw the behavior of the quantity of water in the tub on the second graph below.

Assume the initial quantity in the tub (at time zero) is 100 liters.

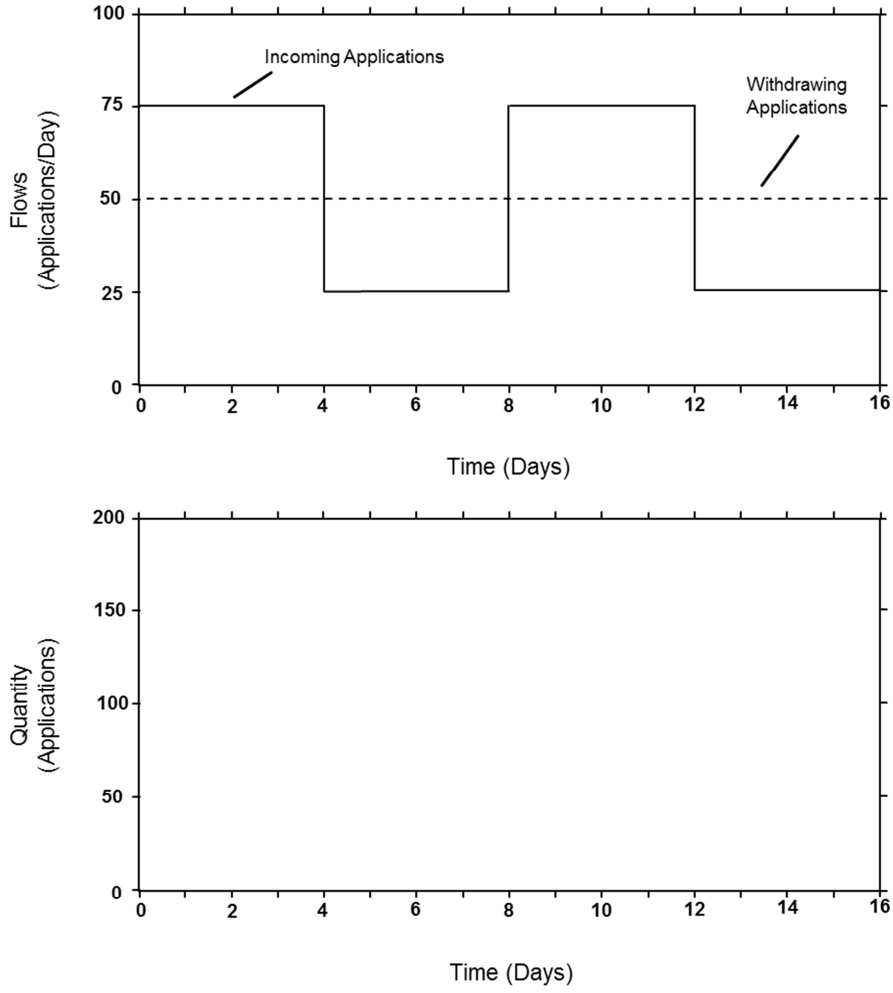


“Bathtub Dynamics” Task I (b)

Consider the pile of online applications of employment shown below. New applications are received at a certain rate on the left. Some applications are withdrawn (rate on the right):

The graph below shows the hypothetical behavior of the incoming and withdrawn applications. From that information, draw the behavior of the quantity of applications on the second graph below.

Assume the initial quantity of applications (at time zero) is 100 applications.

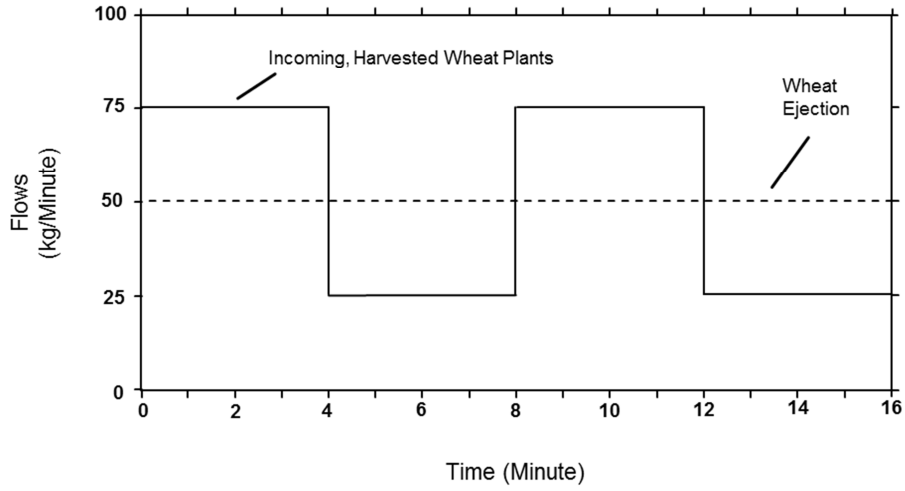


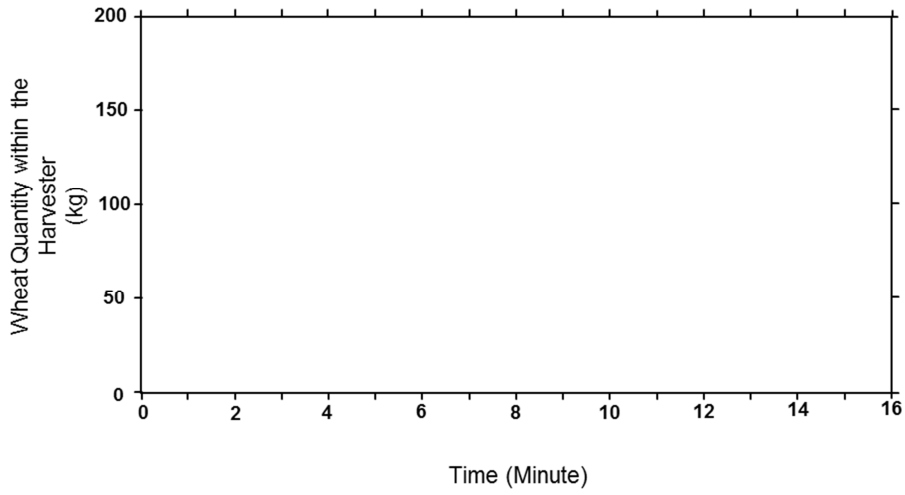
“Bathtub Dynamics” Task I (c)

Consider the harvest of wheat shown below. The harvester harvests wheat plants, processes them and ejects the wheat on the accompanying trailer:

The graph below shows the hypothetical behavior of the incoming, harvested wheat plants into the harvester and the wheat ejection out of it. From that information, draw the behavior of the wheat quantity within the harvester in the second graph below.

Assume the initial quantity of wheat in the harvester (at time zero) is 100 kg.





"Bathtub Dynamics" Task II

Consider the cash balance of a company. Receipts flow in to the balance at a certain rate, and expenditures flow out at another rate:

The graph below shows the hypothetical behavior of the receipts and expenditures. From that information, draw the behavior of the firm's cash balance on the second graph below.

Assume the initial cash balance (at time zero) is 100 €.