Bathtub Dynamics Revisited: Does Educational Background Matter?

Florian Kapmeier[1*], Roland Maximilian Happach[2], and Meike Tilebein[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 2014

Abstract

In prior studies, people's poor performance in dealing with basic systems concepts has been ascribed to different causes. While results indicate that, among other things, domain specific experience and familiarity with the problem context play a role in this stock-flow-(SF-)performance, this has not yet been fully clarified. In this article, we present an experiment that examines the role of educational background in SF-performance. We hypothesize that SF-performance increases when the problem context is embedded in the problem solver's knowledge domain, indicated by educational background. Using the square wave pattern and the sawtooth pattern tasks from the initial study by Booth Sweeney and Sterman (2000), we design two additional cover stories for the former, the Vehicle story from the engineering domain and the Application story from the business domain, next to the original Bathtub story. We then test the three sets of questions on business students. Results mainly support our hypothesis. Interestingly, participants even do better on a more complex behavioral pattern from their knowledge domain than on a simpler pattern from more distant domains. Although these findings have to be confirmed by further studies, they contribute both to the methodology of future surveys and the context familiarity discussion.

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

Introduction

People's understanding of basic systems' structures and their behavior is poor, as indicated by several studies, including the initial survey concerning basic stock-flow-(SF-)relationships by Booth Sweeney and Sterman (2000). The authors asked highly educated subjects to participate in a 10-minute paper-based exercise consisting of two tasks, drawn from a pool of different tasks and cover stories, and a few paragraphs explaining them. The participants were asked to respond by either drawing or interpreting a graph. The results of the survey indicate that the participants have surprisingly poor understanding of basic systems concepts like stock and flow relationships and time delays. This misunderstanding is referred to as SF-failure (Cronin et al. 2009).

Motivated by the initial study and based on the therein described cover stories (Bathtub, Cash flow, Department store, and Manufacturing), scholars worldwide conducted similar surveys concerning SF-performance. While overall performance was equally poor, many of these surveys are not directly comparable because tasks were altered and participant groups were not alike. Differences in tasks include translations (Kapmeier and Zahn 2001; Kainz and Ossimitz 2002; Ossimitz 2002; Kapmeier 2004), replacement of tasks (Sterman and Booth Sweeney 2002; Ossimitz 2002; Moxnes and Jensen 2009; Brunstein et al. 2010) and altered representations of the tasks (Cronin et al. 2009; Fischer and Degen 2012; Gonzalez and Wong 2012). Participant heterogeneity refers to e.g. different age and various educational levels. Participants include high school students (Fisher 2002; Moxnes and Jensen 2009), undergraduates, graduates with and without work experience, MBA students (Kapmeier and Zahn 2001, Fisher 2002; Heinbokel and Potash 2002; Lyneis and Lyneis 2002; Kubanek 2002; Quaden and Ticotsky 2002; Zaraza 2002; Kapmeier 2004), and Ph.D. students (Booth Sweeney and Sterman 2000; Ossimitz 2002). Furthermore, participants have diverse educational backgrounds. The variety of educational backgrounds covers students majoring in business and management (Booth Sweeney and Sterman 2000; Kapmeier 2004; Cronin et al. 2009), engineering (Booth Sweeney and Sterman 2000; Kapmeier 2004), environmental studies (Ossimitz 2002), medicine, biology (Brunstein et al. 2010) and humanities (Booth Sweeney and Sterman 2000; Kapmeier 2004; Brunstein et al. 2010). As stated above, despite altered tasks and heterogeneity of participants, the results generally support findings of the initial study (Booth Sweeney and Sterman 2000).

Scholars propose different underlying reasons for SF-failure. We group them into three categories: (1) systems thinking skills, (2) visualization, and (3) domain specific experience. First, research on systems thinking skills analyzes the participants' problem-solving strategies and asks whether the dynamics of stocks and flows are correctly captured. For instance, Booth Sweeney and Sterman (2000) conjecture the use of an intuitive, yet wrong, heuristic that matches a system's behavior to the shape of its input. Subsequent studies support this presumption and show that participants indeed assume a positive correlation between an input of a system and its behavior (Sterman and Booth Sweeney 2002; Cronin et al. 2009; Sterman 2010; Korzilius et al. 2014). This problem-solving strategy is coined pattern matching (Booth Sweeney and Sterman 2000) or "correlation heuristic" (Cronin et al. 2009: 117). In addition, studies examine whether prior systems thinking training affects SF-performance. Initially, Booth Sweeney and Sterman (2000) find that prior beer distribution game (Sterman 1989) experience enhances performance in the Manufacturing cover story describing one of the tasks of the initial questionnaire. They argue that subjects who had played the beer distribution game may have learned about systems concepts. Similarly, Pala and Vennix (2005) state that performance increases in some of the tasks after system dynamics training. Yet, they also found that performance remains poor or even decreases in other tasks (Pala and Vennix 2005). Their first findings are supported by Sterman (2010), who shows that prior system

dynamics education improves performance in stock and flow tasks. However, he points out that the correlation heuristic still persists as a problem-solving approach for some of the participants (Sterman 2010). In addition to research on the effect of general systems thinking training, Gonzalez and Wong (2012) try to implement prior systems thinking skills by providing participants with analogical solutions. They find that comparisons of a stock and flow task to an analogical problem increases performance especially when surface similarity, the mere appearance between objects, is given (Gonzalez and Wong 2012). By providing analogical solutions, the authors show that comparing two problems may increase SF-performance. Yet, it still needs to be further explored whether providing solutions adds to increasing systems thinking skills. In this particular study, participants did not receive any feedback on their performance. Moreover, it remains unclear whether they successfully transferred the solution on the task and understood the dynamics of the problem described.

Second, research on the effect of visualization on SF-performance shows that participants often have difficulties in reading and interpreting graphs and perform better when being confronted with numerical data in tables (Kainz and Ossimitz 2002). According to Sterman (2002), a reason for low SF-performance may not be the failure of systems thinking but "[p]erhaps the reason people do poorly on these bathtub dynamics is not they don't understand stocks and flows, but that they can't read graphs, or can't do the arithmetic" (p. 507). Other scholars try to change the presentation of the stock and flow tasks by setting up laboratory experiments (Größler and Strohhecker 2012) or representing the graphs in more common illustrations (Cronin et al. 2009; Schwarz et al. 2013; Sedlmeier et al. 2014). These changes in visualization do not significantly affect the performance. In contrast, Veldhuis and Korzilius (2012) find that participants' spatial ability, i.e. the ability to mentally visualize problems, has a positive effect on the performance in stock and flow tasks.

Third, scholars analyze the impact of prior knowledge in a relevant field on performance, like domain specific experience, for example. As mentioned above, Booth Sweeney and Sterman (2000) show that playing the beer distribution game positively affects performance in the Manufacturing task. In addition to the interpretation above they reason that subjects might remember the oscillating supply chain behavior in the beer distribution game after an exogenous step input – without having gained insights about basic systems concepts (Booth Sweeney and Sterman 2000). In other words, subjects use prior knowledge and transfer their domain specific experience to solve a similar task from a familiar field. This is supported by findings that the prior field of study has significant impact on performance (Booth Sweeney and Sterman 2000; Kapmeier 2004). In other words, educational background might matter. These findings indicate that domain specific knowledge, i.e. more background knowledge of and familiarity with a task may affect participants' performance. To investigate this further, Cronin and Gonzalez (2007) and Cronin et al. (2009) change the tasks' contexts to make them intuitively more accessible for participants. Yet, the results do not support their expectations that unfamiliarity with a task's domain contributes to SF-failure (Cronin and Gonzalez 2007; Cronin et al. 2009). In a similar way, Brunstein et al. (2010) analyze the effect of domain experience on SF-performance by designing a number of domain specific cover stories for medical students (Brunstein et al. 2010). Their study shows that domain specific experience enhances performance in some of the tasks, but remains poor in others.

To sum up, none of the three categories can fully explain SF-failure. While there is general agreement on SF-failure as such and on some reasons for the mediocre SF-performance, we also identify academic voids. Researchers find statistically significant differences in performance due to prior academic background especially in the area of domain specific experience (Booth Sweeney and Sterman 2000; Kapmeier 2004; Brunstein et al. 2010) but it stays unclear how domain specific experience affects SF-performance.

With this article, we analyze the effect of domain specific experience. We take educational background as an indicator for domain specific knowledge and link it to problem solving performance on tasks from different knowledge domains. In particular, we investigate business students' SF-performance on different tasks from the domains of business, engineering, and daily life. This article builds on three studies (Cronin and Gonzalez 2007; Cronin et al. 2009; Brunstein et al. 2010) in order to examine the role of prior domain specific knowledge and experience. We expect that performance in SF-tasks increases when the problem context is embedded in the participants' educational background. Our results mainly support our hypothesis. Interestingly, participants even do better on a more complex behavioral pattern from their knowledge domain than on a simpler pattern from a foreign domain.

The article is organized as follows. In the next section, we describe our hypothesis about domain specific knowledge and experience and reason why we expect educational background to be important. In the third section, we lay out the experiment, i.e. the method and the participants. We then present and discuss the results. In the last section we discuss limitations of our study and point to future research paths.

Relevance of educational background

In the following, we first present different findings from the general literature on problem solving. We then reflect on the importance of educational background in SF-performance. We also show that recent studies within the field of system dynamics and SF-failure do not entirely clarify the role of educational background. Based on this, we elaborate our hypothesis.

The role of educational background for problem solving

Scholarly discussions relate problem-solving performance to different dimensions, including (1) interest and motivation, and (2) prior domain specific knowledge and experience. Furthermore, research on problem solving points out that the problem-solving process can be subdivided into two distinct phases: the comprehension phase, or problem representation phase, and the solution phase (Chi et al. 1981; Kotovsky et al. 1985; Gick 1986; Koedinger and Nathan 2004). The comprehension phase precedes the solution phase, thus making comprehension essential for problem solving. In the following, we discuss the link between the two dimensions mentioned above and comprehension.

First, literature on the dimension of interest and motivation shows that interest invokes deeper comprehension (Tobias 1994) and is required for successful problem solving (Mayer 1998). Motivation increases the efforts to fully understand a problem and therefore leads to more effective problem solving (Chi et al. 1989). Motivation can be classified into intrinsic and extrinsic motivation (Ryan and Deci 2000). Extrinsic motivation, on the one hand, is triggered by externally provided rewards (e.g. money, grades, status etc.) (Deci 1972). The effect of extrinsic motivation stays unclear, however. Though there is evidence that incentives and reward may improve motivation and thus performance, some experiments show that incentives lead to decreased performance (Cramerer and Hogarth 1999). Intrinsic motivation, on the other hand, is based on inherent satisfaction (e.g. fun, challenge, eagerness, etc.) (Ryan and Deci 2000). The effect of intrinsic motivation on performance has yet not been fully clarified. Though there is considerable research on the effects of intrinsic motivation on time spent for voluntarily solving tasks and understanding (Deci 1972; Ryan and Deci 2000), we did not find any empirical studies focusing on the direct link between intrinsic motivation and performance.

Second, the dimension of prior domain specific knowledge and experience has been the focus of several studies and experiments. Prior domain specific knowledge (or domain knowledge) is defined as knowledge that "deals with familiarity with general information in an area" (Tobias 1994, p. 39) and "includes conceptual and factual knowledge and procedures associated with a particular domain" (Schraagen 1993, p. 287). It contributes to the understanding of a novel problem by several means. Some scholars find that prior domain specific knowledge provides analogous problems that a person has solved (Gick and Holyoak 1980; Gick 1986; Koedinger and Nathan 2004). In that way, problem solvers may apply familiar problem-solving procedures to the novel problem. Other scholars find that prior domain specific knowledge helps structuring problems (Chi et al. 1981; Schraagen 1993), leading to understanding and applying appropriate approaches for problem solving (Chi et al. 1981). Further studies argue that prior domain specific knowledge provides additional implicit knowledge which helps problem solvers to focus on deep-level information of the problem description (Chi et al. 1981; Gobbo and Chi 1986). It thereby provides substance for reasoning (Schraagen 1993). Kotovsky et al. (1985) show that different cover stories enhance or decrease performance when solving Tower of Hanoi problems. It is argued that familiarity with a topic plays an important role in problem-solving performance. Additionally, embedding problems into an existing knowledge base of a problem solver reminds her of successful prior problem solving (Bandura 1977), which increases self-efficacy yielding better performance (Bandura 1993; Mayer 1998). There is, by contrast, other empirical evidence that such confidence does not necessarily improve performance (Oskamp 1965).

Regarding methodology, the influence of motivation and prior domain specific knowledge on performance has been revealed in experimental settings. Participants in these experiments are either children (Gobbo and Chi 1986; Stern and Lehrndorfer 1992) or students (Gick and Holyoak 1980; Chi et al. 1981; Chi et al. 1989; Schraagen 1993; Koedinger and Nathan 2004). The different degrees of knowledge (novice, intermediate, experts) are either based on educational levels, such as undergraduates, graduates or Ph.D. students within one major (Chi et al. 1981; Schraagen 1993), on grades and pretests (Gobbo and Chi 1986; Chi et al. 1989), or on educational background, such as the study major (Schraagen 1993), or whether or not a specific class had been taken (Koedinger and Nathan 2004).

To sum up, this line of research provides clear hints that educational background plays a role in problem-solving performance in general.

The role of educational background in SF-failure

The laid out dimensions of motivation and domain specific knowledge are not only relevant for general problem solving but also for solving SF-tasks. While some of the studies concerning SF-failure try to isolate and eliminate these effects, other studies try to focus their investigation specifically on these effects. Studies trying to eliminate the effect of motivation and domain specific knowledge and experience include Booth Sweeney and Sterman (2000), Moxnes (2000), Sterman and Booth Sweeney (2002) among others. In their initial survey, Booth Sweeney and Sterman (2000) intently chose the bathtub and cash flow context because "[b]oth cover stories describe everyday contexts quite familiar to the subjects" (Booth Sweeney and Sterman 2000, p. 252). Moxnes (2000), referring to his cover story on the interplay between reindeer herds and lichen, states "I did not ask the participants about their knowledge of the individual parts of the system. However, the Saami reindeer herders are able to correctly describe how numerous species of lichen grow and are eaten by reindeer. The inexperienced participants obtained sufficient information to figure out the structure of the model" (Moxnes 2000, p. 340). Other studies argue accordingly that intuitive usual problems should be common to the

participants (Booth Sweeney and Sterman 2000; Cronin and Gonzalez 2007; Cronin et al. 2009). However, in a later article, Sterman and Booth Sweeney (2002) question whether the tasks' contexts are common enough to the participants or additional information is needed. Additional information is likely to increase understanding of the problem and readers' interest, motivation, and thus effort to solve the problem at hand.

Other studies try to focus their investigation specifically on the effects of problem context and the participants' familiarity with it. For instance, Cronin and Gonzalez (2007) change the cover story of the Department store task into a more intuitive context – a private bank account – by assuming that everyone is familiar with it because of personal experience. Thus, in contrast to Booth Sweeney and Sterman (2000), this cover story embeds the problem within personal experience because it is a private bank account (Cronin and Gonzalez 2007) while the initial study presents a company's bank account in its Cash flow task (Booth Sweeney and Sterman 2000). The explanation is similar to prior studies: authors assume that tasks embedded in usual or daily or ubiquitous situations are familiar to people, be it a bathtub (Booth Sweeney and Sterman 2000), a bank account (Cronin and Gonzalez 2007), distance between cars (Cronin et al. 2009), or reindeer management when asking herders (Moxnes 2000). We believe these assumptions are justifiable when investigating people's general problem-solving performance. However, people do not usually approach everyday problems with an analytical mindset, e.g. the regulation of a bathtub's level or keeping the distance to a proceeding car. As SF-failure surveys do require an analytic approach, we suggest embedding their tasks in people's professional context, thus building on their educational background, and where analytical approaches are more usual.

Only Brunstein et al. (2010) focus on domain specific problem contexts in SF-tasks (Brunstein et al. 2010). The authors examine the effect of the cover story on SF-performance of medical students by designing six cover stories, five of which are embedded in the medical domain. They show that medical students outperform other students in some of the tasks, including general tasks from the initial study. However, in some of the domain specific tasks all participants performed equally poor. According to Brunstein et al. (2010), domain specific knowledge may not have a strong effect on SF-performance. However, they point out several weaknesses of their study. First, the method is not consistent – medical students filled out an online test and other students used a paper test. Second, participants' average domain experience amounts to only 1.5 years. Third, the experiment was only conducted within the medical domain (Brunstein et al. 2010). In addition to these drawbacks, we argue that the medical cover stories in this study may differ in difficulty of understanding since some domain specific task performance was significantly worse than others. It may be that some medicine specific task description or technical terms stem from specific courses the participants did not take.

The tasks of the initial study (Booth Sweeney and Sterman 2000) were meant to be general and to represent a dynamic system of a usual situation, like e.g. filling a bathtub. We assume that this simple problem description does not offer an easy access to problem solving. Instead, it might even seem to be more difficult to solve the task. Participants do not only face a new and abstract concept (systems thinking), but also a novel problem which does not relate to their prior education (the bathtub problem context). Thus, we suggest that the 'familiar' tasks developed for research on SF-performance are relatively abstract considering that people rarely think about stocks, flows, and time delays. In addition, even though the questions relate to everyday situations, they do not activate participants' problem-solving capabilities from their educational background. Therefore, we suggest creating tasks that link the dynamics to the participants' prior knowledge. In line with research on domain specific knowledge and experience, we assume that the educational background (current/prior field of

study) is a strong indicator for prior domain specific knowledge (Chi et al. 1981; Chi et al. 1989; Schraagen 1993; Chi 2006; Brunstein et al. 2010). Hence, we argue:

Hypothesis: Stock and flow performance increases when the problem context is embedded in the educational background of the problem solver.

The hypothesis leads to two expectations:

- Firstly, we expect that domain specific tasks with simple behavioral pattern are significantly better solved than tasks that are distant from the problem solver's knowledge domain.
- Secondly, we expect that domain specific cover stories enhance SF-performance even when the underlying behavioral pattern is more complicated. In other words, subjects might perform better in a more complicated systems thinking task which is embedded in their educational background than in an easier systems thinking task of which the problem context is rather distant from their educational background.

Method

Method and Solution

In this section we present the method we used to conduct the research. In order to guarantee comparability between earlier research (Booth Sweeney and Sterman 2000; Kapmeier and Zahn 2001, Kainz and Ossimitz 2002; Ossimitz 2002; Sterman 2002; Sterman and Booth Sweeney 2002; Fisher 2002; Heinbokel and Potash 2002; Kubanek 2002; Lyneis and Lyneis 2002; Quaden and Ticotsky 2002; Zaraza 2002), and our results, the structure and the content of the original tasks were mainly retained, as described in the following.

We use two of the four classic 'bathtub dynamics' tasks, the 'Bathtub' and the 'Cash flow' cover stories with their underlying square wave and sawtooth patterns for the inflow (Booth Sweeney and Sterman 2000) as a basis. To test our hypothesis we design different cover stories to replace the original Bathtub cover story, leading to three sets of questions and groups: a control group and two experimental groups. Each set of questions consists of two tasks, and their sequence goes along with Booth Sweeney and Sterman's (2000) study: in all tasks, simple patterns of inflows and outflows to a single stock are given with the outflows being constant. Only the inflow patterns differ: the first task entails the square wave pattern and the second task the sawtooth pattern (see Figure 1) for the inflow.

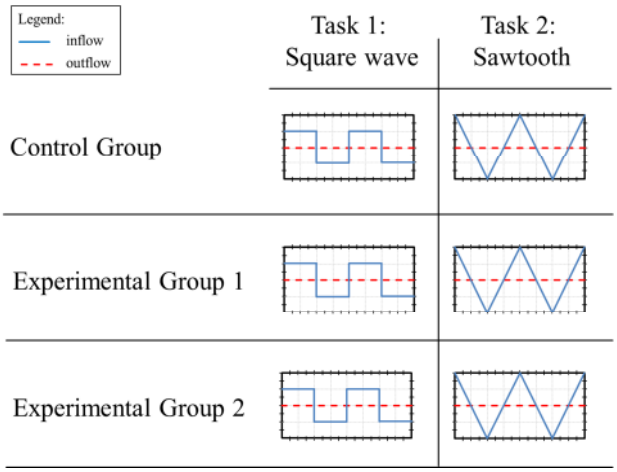


Figure 1: The sequence of behavioral patterns for inflow (solid blue line) and outflow (dotted red line) is the same for all three sets of questions

The behavioral patterns for the inflow and the outflow go along with different cover stories. Cover stories for the Control Group are similar to the classic cover stories described by Booth Sweeney and Sterman (2000). As can be seen from the Figure 2, for Experimental Groups 1 and 2 we developed new cover stories: Experimental Group 1 entails a story that is business related, whereas Experimental Group 2 is engineering related. For the business-related cover story we consider online job applications of employment (the stock). Figure 2 shows the visualizations of the the tasks for the Control and Experimental Groups. They are depicted in the respective set of questions handed out to the participants. Note the similar way of representation: a stock is filled by an inflow (on the left-hand side of the stock) and depleted by an outflow (on the right-hand side of the stock): for the business related task (Experimental Group 1), new applications are received at a certain rate and some applications are withdrawn (rate on the right). For the engineering related task (Experimental Group 2) we consider the speed of a car (the stock). Speed increases by the car's acceleration (the inflow), and declines through braking (the outflow).

For business students, the Application story is an example of a situation within HR Management. A basic business study program class usually covers this topic in the first two semesters of a Bachelor or Master program. Yet, the Vehicle cover story falls more into the specific knowledge domain of engineering students.

In our experiment, we include only one domain specific cover story and not several, like done by Brunstein et al. (2010), for example. Doing so, we minimize the risk that the domain specific cover stories differ in complexity and difficulty. Further, we do not argue that the situations described are common and daily situations for the participants (cf. Booth Sweeney and Sterman 2000, Sterman and Booth Sweeney 2002). We specifically intend to describe situations that are embedded in the participants' educational background.

Further, all sets of questions share the same second cover story: the sawtooth pattern comes along with the Cash flow story from the initial study (Booth Sweeney and Sterman 2000). For the Control Group, the square wave pattern for the first task is embedded in the original Bathtub cover story. Thus the Control Group receives the classic sequence of Bathtub and Cash flow cover stories as described in Booth Sweeney and Sterman (2000). Handing out this questionnaire to the students enables us to compare the students' performance with the many SF-performance studies carried out earlier. This Control Group helps to ensure that interpretations drawn from results with the experimental groups are valid.

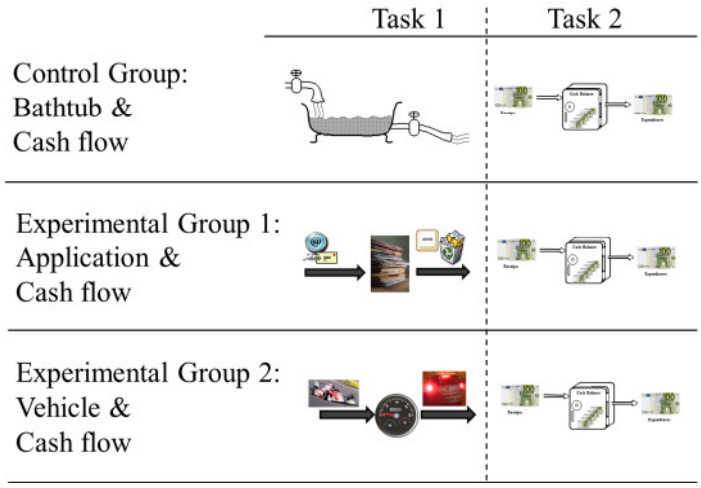


Figure 2: The cover stories: Bathtub & Cash flow, Application & Cash flow, and Vehicle & Cash flow with their respective visualizations for the sets of questions containing task 1 and task 2 (stocks with their inflows and outflows)

In the following section we present the subjects and the procedure that we followed to conduct the survey.

Subjects and Procedure

The tasks were handed out to a group of Business students at the ESB Business School at Reutlingen University, Germany. The students were enrolled in the 'International Business' (IB) Bachelor of Science study program at ESB. It is a seven semester Bachelor program taught in English with roughly 50% students originating from out of Germany. The 39 participants of the study were enrolled in the 5th-semester courses 'Business Simulation' in the 'Marketing & Strategy Major' or 'Consolidated Financial Statements' in the 'Finance & Accounting Major'. Both are mandatory classes within their respective major fields. The courses were offered in the summer term 2014.

All IB students had the same teaching content until the fifth semester. With the beginning of choosing the Major subjects, teaching content starts to differ. To keep differences in the university education background in both groups as low as possible, the tasks were distributed to both student groups on their first day of class in that particular semester.

As with the original study conducted at MIT(Booth Sweeney and Sterman 2000), students were told that the purpose of the tasks was to gain insight into people's understanding of systems thinking before they were introduced to System Dynamics. They were also informed that they participate in a longitudinal analysis conducted by the MIT's System Dynamics Group. As before (Kapmeier 2004), it was stressed that the participants' performance on the tasks would not influence the students' grades. Students were not being paid. The group of participants was arbitrarily subdivided into a control group and two experimental groups, each of which was handed out their respective set of questions. The participants had 10 minutes to work on the task after filling out the cover page in which we asked the participants for the background information needed to carry out the statistical analysis. After 5 minutes we pointed out to the students that they should turn to the next question. Yet, it was not checked whether they had actually done it.

As can be seen from the Table 1, the background data sheet asked for information about the participants' age, gender, and current degree program, 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 female students (77%). As can be seen from the Table 2, nearly all subjects were in their 5th semester, despite eight students, who were international exchange students. The large majority of the group (87%) were younger than 24. As the IB program is characterized by admitting a considerable part of the student body from outside of Germany, only a minority were German native speakers (33%). Still, the majority originated from Europe (71%). Some students came from Asia and the Middle East (21%) and few from North (5%) and South (3%) America. Due to this vast variety of countries of origin, English, the study program and task language, was the first language only for a minority of students (15%). None of the participants had played the beer distribution game before.

	ESB Summer 2014 BSc International Business			
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	**39**	**14**	**13**	**12**
Age				
19-24	34	11	13	10
25-30	5	3	0	2
31-35	0	0	0	0
36 and up	0	0	0	0
Gender				
Male	13	3	5	5
Female	24	9	8	7
Student Status				
IB program	31	12	10	9
IB exchange student	8	2	3	3
Prior Field of Study				
Business/Management	0	0	0	0
Engineering	0	0	0	0
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
BS	0	0	0	0
MA, MS, Diplom	0	0	0	0
Ph.D.	0	0	0	0
High School	39	14	13	12
BE. JD, BBA, MD, CPA	0	0	0	0
Current Field of Study				
Business/Management	39	14	13	12
Engineering	0	0	0	0
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
Region of Origin				
North America (Aus. + NZ)	2	1	1	0
Europe	28	9	10	9
Asia and Middle East	8	4	1	3
Latin and South America	1	0	1	0
Africa	0	0	0	0
First language				
German	13	5	2	6
English	6	2	2	2
Other	20	7	9	4
Teaching Language				
English	39	14	13	12
German	0	0	0	0
Beer Game Experience				
Played before	0	0	0	0
Have not played	39	14	13	12

Table 1: Subject demographics – ESB Business School, 'BSc International Business' study program

Results

In this section 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 2 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.

The overall result was poor with the average performance resulting in roughly 50% correct answers. In the initial study by Booth Sweeney and Sterman (2000), 77% answered correctly, thus 27 percentage points higher than performance by ESB students.

Criterion	Square Wave pattern Average BSc International Business	Square wave pattern *Bathtub* cover story (Control Group) BSc International Business
1. When the inflow exceeds the outflow, the stock is rising.	0.52	0.57
2. When the outflow exceeds the inflow, the stocks is falling.	0.57	0.50
3. The stock should not show any discontinuous jumps (it is continuous).	0.70	0.57
4. The peaks and troughs of the stock occur when the net flow crosses zero (i.e., at t= 4,8,12,16)	0.54	0.57
5. During each segment the net flow is constant so the stock must be rising (falling) linearly.	0.52	0.50
6. The slope of the stock during each segment is the net rate (i.e., +/-25 units/time period).	0.31	0.28
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.	0.29	0.28
Mean for all items	0.49	0.47

		correlation with Control Group				correlation with Control Group		
	Square wave pattern *Application* cover story BSc International Business	X^2	*Cramer's V*	*p*	Square wave pattern *Vehicle* cover story BSc International Business	X^2	*Cramer's V*	*p*
1.	0.75	0.114	0.098	0.735	0.23	2.400	0.433	0.12
2.	0.83	2.400	0.447	0.121	0.39	2.236	0.415	0.135
3.	0.83	1.714	0.378	0.190	0.69	0.325	0.158	0.569
4.	0.75	0.114	0.098	0.735	0.31	0.442	0.184	0.506
5.	0.83	2.400	0.447	0.121	0.23	3.343	0.507	0.067
6.	0.58	0.686	0.239	0.408	0.08	2.438	0.433	0.118
7.	0.50	1.500	0.354	0.221	0.08	2.438	0.433	0.118
	0.72				0.29			

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

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 ESB students' performance on the Bathtub cover story and the square wave pattern to that of MIT's students (Booth Sweeney and Sterman 2000) we notice that performance was much poorer with only 47% being correct (MIT: 83%). Thus, the average performance rates 36 percentage points lower. However, our result goes along with other studies, like Ossimitz' (2002) study, who observed an average performance of 42% with 154 participants. Just like MIT students, the ESB students did best on stating correctly that the stock does not show any discontinuous jumps (57%, compared to 96% with MIT). Considerable fewer students stated correctly the rising (57%, MIT: 87%) and falling (50%, MIT: 86%) of the stock at the appropriate times. Performance was also poor when stating correctly that the slope of the stock during each segment is the net rate (28%, MIT: 73%), and only 28% (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 criterion 3 being the best and criteria 6 and 7 the worst.

Interestingly, while the best and worst individual coding criteria are similar for the Control Group and the Experimental Groups, the result differs when analyzing the actual magnitude of performance. 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 (72%) than performance with the Vehicle story (29%). 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 (83% Application cover story, and 69% Vehicle cover story). The majority of the students from Experimental Group 1 also stated correctly the rising (75%) and falling (83%) of the stock at the appropriate times. This is considerably higher than with the classic Bathtub story (stock rising: 57% and stock falling: 50%). Yet, students with the Vehicle cover story had difficulties indicating that the stock is rising (23%) or falling (39%). The largest difference can be seen when comparing the performance on the correct slope of the stock (58% with the Application story and only 8% with the Vehicle story) and the correct quantity added to the stock each segment (50%, 8%). These are also the two criteria with the worst performance for the two cover stories: this means that for the Application cover story, half of the participants gave wrong answers, and everybody except for one student stated this incorrectly for the Vehicle story.

Sterman (2002) observed pattern matching as an often occurring error in the MIT group's results. Likewise, 29% of the IB students working on the Bathtub story and 23% of the IB students working on the Vehicle story also matched the pattern for the stock to the inflow. This was the most typical error of these two student groups. Yet, only one individual replicated the inflow pattern for the stock in the Application cover story (8%).

As observed before (Kapmeier 2004) it can also be stated here that the first criteria correlate highly (Bathtub: Pearson's R=0.866, Application: Pearson's R=0. 775, and Vehicle: Pearson's R=0. 693) and significantly (Bathtub: p=0.000, Application: p=0.003, and Vehicle: p=0.009) 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 3, average performance of the IB students on the Cash flow task was poor (38%). There are only small deviations from average when looking at the performance, in contrast to the considerable performance differences on the previous square wave pattern tasks: students who had previously worked on the Bathtub cover

story perform worst (Control Group: 35%), followed by Experimental Group 1 with the Application cover story (38%), and Experimental Group 2 with the Vehicle cover story (41%). When compared to previous studies, in which 45% (Kapmeier 2004), 51% (Booth Sweeney and Sterman 2000), and 48% (Ossimitz 2002) correctly answered the Cash flow task with the underlying sawtooth pattern, performance is slightly lower. Note that while average performance on this task is lower than with the square wave pattern task, there is one exception: Experimental Group 2 performed considerably better in the second task (41%) than in the previous vehicle task (29%).

	Criterion	Sawtooth pattern Average BSc International Business	Sawtooth pattern Previous cover story: Bathtub (Control Group) BSc International Business
1.	When the inflow exceeds the outflow, the stock is rising.	0.49	0.43
2.	When the outflow exceeds the inflow, the stock is falling.	0.45	0.43
3.	The stock should not show any discontinuous jumps (it is piecewise continuous).	0.88	0.71
4.	The peaks and troughs of the stock occur when the net flow crosses zero (i.e., t = 2, 6, 10, 14).	0.49	0.43
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).	0.15	0.14
6.	The slope of the stock when the net rate is at its maximum is 50 units/period (t = 0, 8, 16).	0.12	0.21
7.	The slope of the stock when the net rate is at its minimum is -50 units/period (t = 4, 12).	0.18	0.29
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.	0.26	0.14
	Mean for all items	0.38	0.35

		correlation with Control Group				correlation with Control Group		
	Sawtooth pattern Previous cover story: Application BSc International	X^2	*Cramer's V*	*p*	Sawtooth pattern Previous cover story: Vehicle BSc International	X^2	*Cramer's V*	*p*
1.	0.50	0.343	0.169	0.558	0.54	0.737	0.238	0.391
2.	0.50	0.343	0.169	0.558	0.41	0.066	0.071	0.797
3.	0.92	2,182	0.426	0.140	1.00	n.a.	n.a.	n.a.
4.	0.50	0.343	0.169	0.558	0.54	0.737	0.238	0.391
5.	0.17	0.480	0.200	0.488	0.15	2176.00	0.409	0.140
6.	0.08	0.364	0.174	0.546	0.08	0.325	0.158	0.569
7.	0.17	0.300	0.158	0.584	0.08	0.481	0.192	0.488
8.	0.17	1,920	0.400	0.166	0.46	2,026	0.395	0.155
	0.38				0.41			

Table 3: Performance on the sawtooth pattern tasks, differentiated according to the previous Bathtub, Application, and Vehicle cover stories – students of the 'BA International Business' study program. The X^2 statistic tests the hypothesis that performance on the two treatment conditions is the same

Around half of the IB students showed correctly that the stock rises (falls) when the inflow exceeds the outflow (and vice versa): 43% (43%) answered correctly when the previous cover story was the Bathtub story, 50% (50%) the Application story, and 54% (41%) the Vehicle story. Further, more or less the same participants from the three groups who succeeded in this criterion also marked the peaks and troughs of the stock at the appropriate points in time (43% of Bathtub group, 50% of the Application group, and 54% of the Vehicle group). A fairly high number of IB students (85%) failed to relate the net rate to the stock (86% Bathtub, 83% Application, 85% Vehicle). This is roughly more than 15 percentage points more than the original MIT group (Booth Sweeney and Sterman 2000). Whereas substantially more participants from the MIT study correctly drew the maximum (52%) and the minimum (51%) slope of the stock, significantly fewer IB students did so (21% and 29% for the Bathtub cover story, 8% and 17% for the Application cover story, and 8% and 8% for the Vehicle cover story). Whereas 100% (Vehicle) of the IB students recognized that there are no discontinuous jumps in the stock in the 'Cash flow' task all but 1 (92%, Application) and 4 students (71%, Bathtub) drew the stock without any discontinuous jumps. Hence, the number approximately equals MIT students' performance (99%). Whereas nearly half (46%) of the IB students who had worked previously on the Vehicle task correctly calculated the maximum and minimum of the stock, only 14% (Bathtub) and 17% (Application) did so. To do so, students simply had to calculate the area of the triangle bound by the net rate. In other words, 54% (Vehicle), 86% (Bathtub), and 83% (Application) of the groups failed to apply graphical integration correctly to this challenge, respectively.

As in the square wave task, pattern matching was one of the most common errors for the groups (45% in the Bathtub group, 40% in the Application group, and 72% in the Vehicle group). 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 and Application: Pearson's R=1.000; Vehicle: Pearson's R=0.857 and p=0.000), and between criteria 1, 2 and 4 (peaks and troughs) (Bathtub and Application: Pearson's R=1.000 for both criteria; Vehicle: Pearson's R=1.000 for 1 and 4 and Pearson's R=0.857 and p=0.000 for 2 and 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 (Pearson's R=0.732 and p=0.004) and no significant correlation in the Application square wave and the sawtooth tasks (Pearson's R=0.192 and p=0.549) and the Vehicle square wave and the sawtooth tasks (Pearson's R=0.507 and p=0.077). Likewise is the correlation between criteria 2 in the respective tasks (Bathtub square wave and Cash flow sawtooth: Pearson's R=-0.732 and p=0.004; Application square wave and Cash flow sawtooth: Pearson's R=0.192 and p=0.549; Vehicle square wave and the Cash flow sawtooth tasks: Pearson's R=0.592 and p=0.033). The first and the latter are correlating. There is no significant correlation between the remaining criteria.

Discussion of Results

For our study, we argue that SF-performance increases when the problem context is embedded in the problem solver's knowledge domain. We derive two expectations from this hypothesis: Firstly, we expect that problem solvers perform significantly better on simple tasks from their specific knowledge domain than tasks

rather distant from their knowledge domain. Secondly, we expect that domain specific cover stories enhance SF-performance even when the underlying behavioral pattern is more complicated.

In its general findings regarding typical mistakes or ranking of performance criteria, for example, the results of our study are in line with prior studies. Specifically, our results are as follows. Results of the square wave pattern task support our first expectation. Participants majoring in business perform best in the task with the Application cover story, a topic which stems from their knowledge domain. The second best results are achieved in the task with the Bathtub cover story. Last, participants performed worst in the task with the Vehicle cover story - this particular context is out of their knowledge domain. Moreover, SF-performance in the task embedded in domain knowledge is considerably higher than in the two other tasks. These differences in performance are statistically significant. Consequently, the results suggest that domain specific tasks with simple behavioral patterns are significantly better solved than tasks from a more distant knowledge domain.

Our results also qualify a ranking in performance. In addition to the results mentioned above, we find a difference between the performance in the Bathtub and Vehicle cover stories. According to these results, people perform significantly better in tasks embedded in a usual everyday problem context than in a task embedded in a different domain specific problem context. We assume that general familiarity and intuition play an important role in SF-performance as suggested by other scholars (Booth Sweeney and Sterman 2000; Cronin and Gonzalez 2007; Cronin et al. 2009). However, we argue that a better comprehension of the problem context and thus better SF-performance is supported when the task is embedded in a subject's domain specific knowledge.

Moreover, we find hints that confirm our second expectation. Subjects might perform relatively better in a more complicated systems thinking task embedded in their educational background than in an easier one not embedded in their educational background. Specifically, subjects with business background perform better when solving the Cash flow task than a task embedded in a more distant domain, such as the tasks with the Bathtub or Vehicle cover story – despite the less complex behavioral pattern in the latter two cases. Our findings show that the second task was solved nearly equally well (or poorly) by all groups. This does not come as a surprise as the participants hold prior knowledge in the domain of business. However, analysis indicates small differences. Experimental Group 2 performed best, Experimental Group 1 performed second best and the Control Group performed worst. In relative terms, performance of Experimental Group 1 drops from high above average in the previous task with the Application cover story to average in the task with the Cash flow cover story, Experimental Group 2 here performs better than on the previous task with the Vehicle cover story, and the Control Group again performs around average.

In general, as stated above, our results go along with results from previous studies. Similar to previous studies (Booth Sweeney and Sterman 2000), our analysis suggests that solving the first task does not influence the performance of the second task: the tasks are independent from each other. One of the reasons for this could be that subjects may start working on the second task before the first task – the opposite way than intended. So, we do not have information about the sequence of solving. This is mainly due to the design of the experimentation set-up. Also, since participants do not get any feedback on their performance of the first task and thus are not trained in system thinking skills, we cannot expect that those who correctly solve the first task will also perform well in the second.

Further, we found remarkable that participants of Experimental Group 2 performed relatively poor on their first task with the Vehicle cover story and yet high on their second task (Cash flow cover story). We conjecture that they just had not been in their knowledge domain when working on their first task. Yet, when

continuing with the (more complex) second sawtooth task, they then dealt with a cover story that was in their knowledge domain. As the problem context of the second task is again embedded in their knowledge domain it supports the subjects to better comprehend the problem. Nevertheless, subjects in the Control Group performed slightly worse on the second task with the Cash flow cover story even though they also returned to a task embedded in their knowledge domain. We conjecture that the time spent per task may shed new light on the difference. Since we do not have information about the time spent per task, it is possible that participants of Experimental Group 2 might have moved on with working on the second Cash flow task early after they had realized that they simply did not grasp the underlying Vehicle storyline of task 1. Accordingly, participants of the Control Group might have spent more time for the Bathtub task than their Experimental Group 2 counterparts spent on the Vehicle Task. They therefore might have had less time to solve the task. We infer that more research is needed in order to support our conjecture and to establish statistically significant results on this expectation.

Finally, results from the experiment reported here also include some drawbacks. Firstly, we conducted our experiment with only 13 participants per group. Therefore, significant results lose power. Secondly, we argue that poor performance in the task with the Vehicle cover story is due to unfamiliarity because it stems from a different domain of knowledge. Alternatively, there might be more reasons for the poor performance. Though we intended to design comparable and equally complex cover stories for the respective knowledge domains, there are slight differences regarding tangibility and countability (see Table 4). Regarding the latter, only the Application cover story deals with countable items which some subjects may find easier to access. Concerning tangibility, the Vehicle cover story is the only one without tangible flows. Velocity is intangible and thus might contribute to explain the poor performance. In contrast, the Bathtub cover story describes water flowing into a bathtub and thus provides a tangible flow. Similarly, subjects might relate online applications to paper applications, turning this storyline into one with tangible flows.

Cover story	Tangibility	Countability
Bathtub	X	
Application	(X)	X
Vehicle		
Cash flow	(X)	X

Table 4: Flows differ in terms of tangibility and countability for the different cover stories

It should be noted that there are also differences in tangibility and countability in the original study (Booth Sweeney and Sterman 2000) as the Cash flow story refers to receipts and payments of monetary flows, and thus tangible and countable flows and the Bathtub story, as noted above, only refers to tangible flows. However, a number of researchers investigated the effect of tangibility on performance and find that it plays a rather minor role (Cronin et al. 2009; Größler and Strohhecker 2012; Schwarz et al. 2013; Sedlmeier et al. 2014).

General Discussion and Future Research

In prior studies, the widespread phenomenon of SF-failure has been ascribed to different causes. There is consensus among scholars that a lack of systems thinking skills belongs to the most prominent of these causes. There is an ongoing discussion, however, on additional causes contributing to SF-failure. While prior studies indicate that, among other things, domain specific experience and familiarity with the problem context might

play a role in SF-performance, this aspect has not yet been fully clarified. This is why we focus on this specific aspect in our study.

In this article, we present an experiment that examines the role of educational background on SF-performance. We argue that SF-performance increases when the problem context is embedded in the problem solver's knowledge domain for which we take educational background as an indicator. Using the square wave and the sawtooth pattern tasks from the initial SF-study by Booth Sweeny and Sterman (2000), the latter task remains unchanged with a Cash flow cover story while we design two additional cover stories for the former task, the Vehicle cover story from the engineering domain and the Application cover story from the domain of business as alternatives for the original Bathtub task. We then test the resulting three different sets of questions on business students. Just like in the initial study, a Control group receives the Bathtub and Cash flow tasks. Experimental Group 1 receives the Application and Cash flow tasks, and Experimental Group 2 has to solve the Vehicle and Cash flow tasks.

While overall performance level is lower than in comparable prior surveys, performance patterns in general are in line with prior studies in terms of e.g. typical errors, ranking of performance criteria, or correlations between performance criteria. Regarding our specific research focus, results confirm our expectation for the simple square wave pattern task: While business students in Experimental Group 1 succeed quite well with the Application cover story from the domain of business, their fellow students in Experimental Group 2 have severe difficulties figuring out the solution of exactly the same task, framed as Vehicle task from the engineering domain. Similarly, the Control Group has difficulties solving the Bathtub task which provides a familiar everyday context, but does not relate to the participants' professional domain experience. Therefore we think that instead of offering an easy access to problem solving, SF-tasks from an everyday context might even be more difficult to solve, as they call for a highly unfamiliar perspective on these familiar tasks.

In the second task with the sawtooth pattern and the Cash flow cover story, the three study groups do quite equally poor. Average performance on this more complex behavioral pattern task is lower than on the simpler square wave pattern task. Experimental Group 2 with the previous Vehicle cover story performs best, followed by Group 1 with the Application cover story and the Control Group with the Bathtub Story. The results of Experimental Group 2 may substantiate our hypothesis: Compared to their performance in the previous Vehicle cover story, their overall performance on the Cash flow task is better, although this task is more complex and thus more difficult to solve. We attribute this to the fact that the Cash flow task, in opposite to the Vehicle task, fits the educational background of the participants.

Although results are promising, there are limits to our study, and our experiment brought up a number of effects we cannot fully explain yet. We group these limits and potential biases in three categories, relating to (1) participants of the study, (2) procedure, and (3) method:

First, the group size of about 13 individuals in each of the three groups in our study is rather small. This calls for extending the experiment to more participants in order to confirm our results. Moreover, all participants were business students. In order to mirror and further confirm our results, it would be interesting to conduct the experiment with comparable engineering students. Our hypothesis suggests that they might perform best on the Vehicle task and have difficulties solving the tasks with cover stories from the field of business.

Second, the procedure of the study conducted accounts for some of its limits. In the Results Discussion Section, we identify differences in time split, i.e. the specific minutes participants choose to spend on each of the two tasks, as one potential cause for the unexpected rank 1 performance of Experimental Group 2 in the Cash

flow task. By separating the task sheets and collecting task 1 before handing out task 2 sheets, for example, we could eliminate this potential bias. Moreover, this procedural adjustment could assure that participants obey to the intended task sequence and solve task 1 first, followed by task 2. While coordination need would be higher in a paper-based test, this adjustment could be easily implemented in a computer-based test.

Finally, we consider two aspects for the methodological limitations. The first relates to the problem of creating comparable cover stories and assessing their familiarity for students as well as their pairwise distance. Furthermore, we identify potential biases concerning the background data sheet of the original questionnaire (Booth Sweeney and Sterman 2000). In particular, these limitations refer to identifying participants' prior domain experience. The original data sheet asks for information about the participants' age, gender, and current degree program, region of origin, first language, highest previous degree, teaching language, and whether they had played the beer distribution game before. As discussed earlier, there is reason to assume that the beer distribution game does not only provide insight in systems thinking, but also in domain specific experience. As none of our participants had played the beer game before, this is not relevant for our results. However, the assessment of prior domain specific experience, crucial for our study, remains incomplete if only based on the information of the participant's current degree program. Participants might have gained prior domain specific experience from other sources like e.g. prior vocational training, internships, or job experience, which the questionnaire does not ask for. Specifically, MBA or Executive MBA students participating in the study (i.e., Booth Sweeney and Sterman 2000) might have gained domain specific knowledge from their functional or industry background the questionnaire does not disclose. For example, for a participant from the Executive MBA program with a BA degree in sociology and a mid-level executive position in the R&D department the datasheet would identify domain specific knowledge only in sociology ('prior field of study'). Hence, her domain specific experience in business and engineering would be ignored. Complementing the background data sheet with the respective questions could further enhance our data. Moreover, in this research path we see opportunities to link our study to team diversity research. Here scholars often take differences in, among others, educational background as proxy for cognitive diversity. While they relate diversity to creativity and innovation, our results suggest a link to SF-failure. Combining the two perspectives might shed additional light on team composition issues.

To sum up, our study has revealed valuable insight on the influence of educational background on SF-performance. It shows that, aside of systems thinking skills, prior domain specific knowledge may have a larger influence on SF-performance than expected. Our preliminary results need further confirmation, though. As sketched above, we therefore plan to broaden the study and extend it to participants from different fields of study. In addition, the assessment of domain-specific prior knowledge needs further investigation.

References

Bandura A. (1977): *Self-efficacy: Towards a unifying theory of behavioural change.* Psychological Review Vol. 84 (2): pp. 191–215.

Bandura A. (1993): *Perceived self-efficacy in cognitive development and functioning.* Educational Psychologist Vol. 28 (2): pp. 117–148.

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

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

Chi M. T. H. (2006): *Two approaches to the study of experts' characteristics.* In: Ericsson KA, Charness N, Feltovich P, Hoffman R (eds) *Cambridge Handbook of Expertise and Expert Performance.* Cambridge University Press, Cambridge, pp 121–130.

Chi M. T. H., Bassok M., Lewis M. W., Reimann P., Glaser R. (1989): *Self-explanation: How students study and use examples in learning to solve problems.* Cognitive Science Vol. 13 (2): pp. 145–182.

Chi M. T. H., Feltovich P. J., Glaser R. (1981): *Categorization and representation of physics problems by experts and novices.* Cognitive Science Vol. 5 (2): pp. 121–152.

Cramerer C. F., Hogarth R. M. (1999): *The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework.* Journal of Risk and Uncertainty Vol. 19 (1-3): pp. 7–42.

Cronin M. A., Gonzalez C. (2007): *Understanding the building blocks of dynamic systems.* System Dynamics Review Vol. 23 (1): pp. 1–17.

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

Deci E. L. (1972): *Intrinsic motivation, extrinsic reinforcement, and inequity.* Journal of Personality and Social Psychology Vol. 22 (1): pp. 113–120.

Fischer H., Degen C. (2012): *Stock-flow failure can be explained by the task format.* Proceedings of the 30th International Confernce of the System Dynamics Society in St. Gallen Switzerland.

Fisher D. (2002): *Student Performance on the Bathtub and Cash Flow Problems.* Proceedings of the 20th International Conference of the System Dynamics Society in Palermo, Italy.

Gick M. L. (1986): *Problem-solving strategies.* Educational Psychologist Vol. 21 (1-2): pp. 99–120.

Gick M. L., Holyoak K. J. (1980): *Analogical Problem Solving.* Cognitive Psychology Vol. 12 (3): pp. 306–355.

Gobbo C., Chi M. T. H. (1986): *How knowledge is structured and used by expert and novice children.* Cognitive Development Vol. 1 (3): pp. 221–237.

Gonzalez C., Wong H.-y. (2012): *Understanding stocks and flows through analogy.* System Dynamics Review Vol. 28 (1): pp. 3–27.

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

Heinbokel J., Potash J. (2002): *Bathtub Dynamics at Vermont Commons School.* Proceedings of the 20th International Conference of the System Dynamics Society in Palermo, Italy.

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., Zahn, E. (2001): *Bathtub Dynamics: Results of a Systems Thinking Inventory at the Universität Stuttgart, Germany*, Working Paper of the Betriebswirtschaftliches Institut der Universität Stuttgart, Lehrstuhl für Planung und Strategisches Management, Stuttgart, now at: http://www.esb-business-school.de/business-school/organisation/professoren-und-dozenten/kapmeier.html

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

Koedinger K. R., Nathan M. J. (2004): *The real story behind the story problem: effects of representations on quantitative reasoning.* The Journal of the Learning Sciences Vol. 13 (2): pp. 129–164.

Korzilius H., Raaijmakers S., Rouwette E., Vennix J. (2014): *Thinking Aloud While Solving a Stock-Flow Task: Surfacing the Correlation Heuristic and Other Reasoning Patterns.* Systems Research and Behavioral Science Vol. 31 (2): pp. 268-279.

Kotovsky K., Hayes J. R., Simon H. A. (1985): *Why are some problems hard? Evidence from Tower of Hanoi.* Cognitive Psychology Vol. 17 (2): pp. 248–294.

Kubanek G. (2002): *Bathtub Dynamics Ottawa.* Proceedings of the 20th International Conference of the System Dynamics Society in Palermo, Italy.

Lyneis J., Lyneis D. (2002): *Bathtub dynamics at WPI.* Proceedings of the 20th International Conference of the System Dynamics Society in Palermo, Italy.

Mayer R. E. (1998): *Cognitive, metacognitive, and motivational aspects of problem solving.* Instructional Science Vol. 26 (1-2): pp. 49–63.

Moxnes E. (2000): *Not only the tragedy of the commons: misperceptions of feedback and policies for sustainable development.* System Dynamics Review Vol. 16 (4): pp. 325–348.

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

Oskamp S. (1965): *Overconfidence in case-study judgments.* Journal of Consulting Psychology Vol. 29 (3): pp. 261–265.

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.

Pala Ö., Vennix J. (2005): *Effect of system dynamics education on systems thinking inventory taks performance.* System Dynamics Review Vol. 21 (2): pp. 147–172.

Quaden R., Ticotsky A. (2002): *Bathtub Dynamics at Carlisle Public Schools.* Proceedings of the 20th International Conference of the System Dynamics Society in Palermo, Italy.

Ryan R. M., Deci E. L. (2000): *Intrinsic and extrinsic motivations: classic definitions and new directions.* Contemporary Educational Psychology Vol. 25 (1): pp. 54–67.

Schraagen J. M. (1993): *How esperts solve a novel problem in experimental design.* Cognitive Science Vol. 17 (2): pp. 285–309.

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: Wassong T, Frischemeier D, Fischer PR, Hochmuth R, Bender P (eds) *Mit Werkzeugen Mathematik und Stochastik lernen – Using Tools for Learning Mathematics and Statistics.* Springer Fachmedien, Wiesbaden, pp 57–70.

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

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

Sterman J. D. (1989): *Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment.* Management Science Vol. 35 (3): pp. 321–339.

Sterman J. D., Booth Sweeney L. (2002): *Cloudy skies: assessing public understanding of global warming.* System Dynamics Review Vol. 18 (2): pp. 207–240.

Stern E., Lehrndorfer A. (1992): *The role of situational context in solving word problems.* Cognitive Development Vol. 7 (2): pp. 259–268.

Tobias S. (1994): *Interest, Prior Knowledge, and Learning.* Review of Educational Research Vol. 64 (1): pp. 37–54.

Veldhuis G. A., Korzilius H. (2012): *Seeing with the mind - The role of spatial ability in inferring dynamics behaviour from graphs and stock and flow diagrams.* Proceedings of the 30th International Confernce of the System Dynamics Society in St. Gallen Switzerland.

Zaraza R. (2002): *Bathtub Dynamics in Portland at SyMFEST.* Proceedings of the 20th International Conference of the System Dynamics Society in Palermo, Italy.

Appendix to:

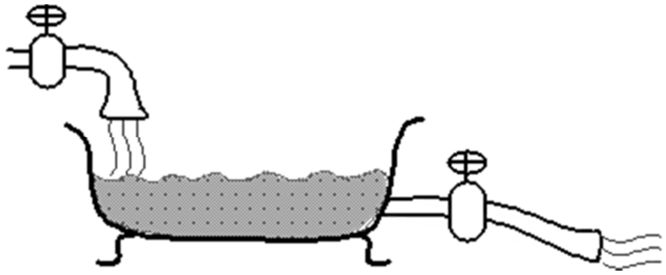
Bathtub Dynamics Revisited:
Does Educational Background Matter?

Note:
The three "Bathtub Dynamics" sets of questions consist of

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

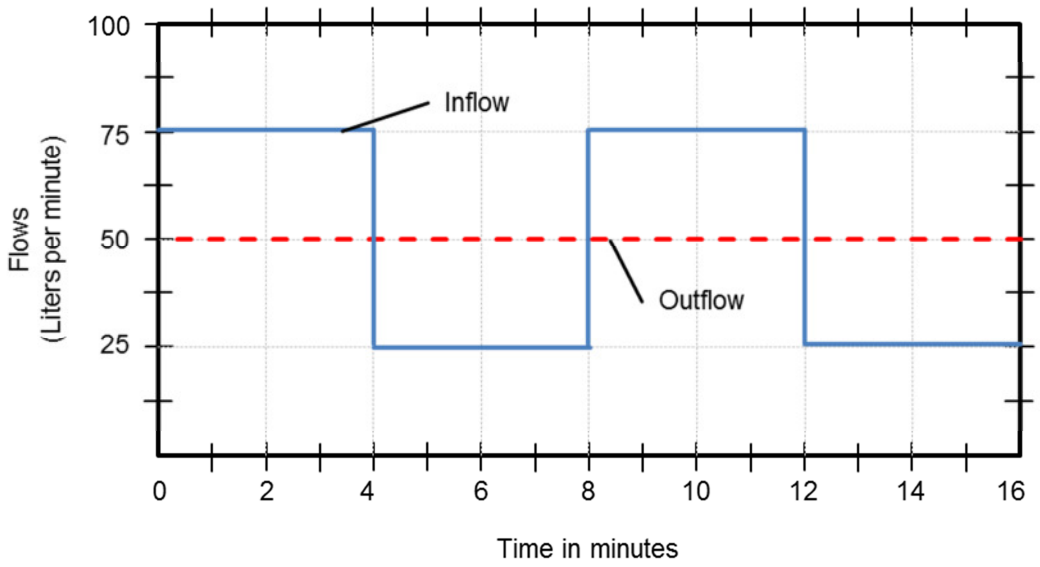
“Bathtub Dynamics” Task I (a)

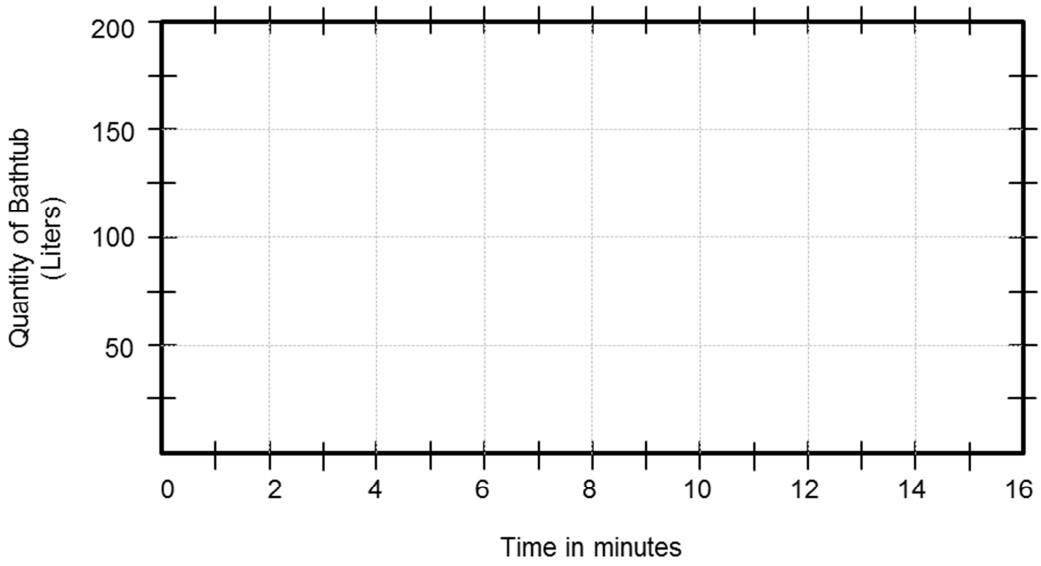
Consider the bathtub shown below. Water flows in at a certain rate on the left, and exits through the drain at another rate (rate on the right):



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.



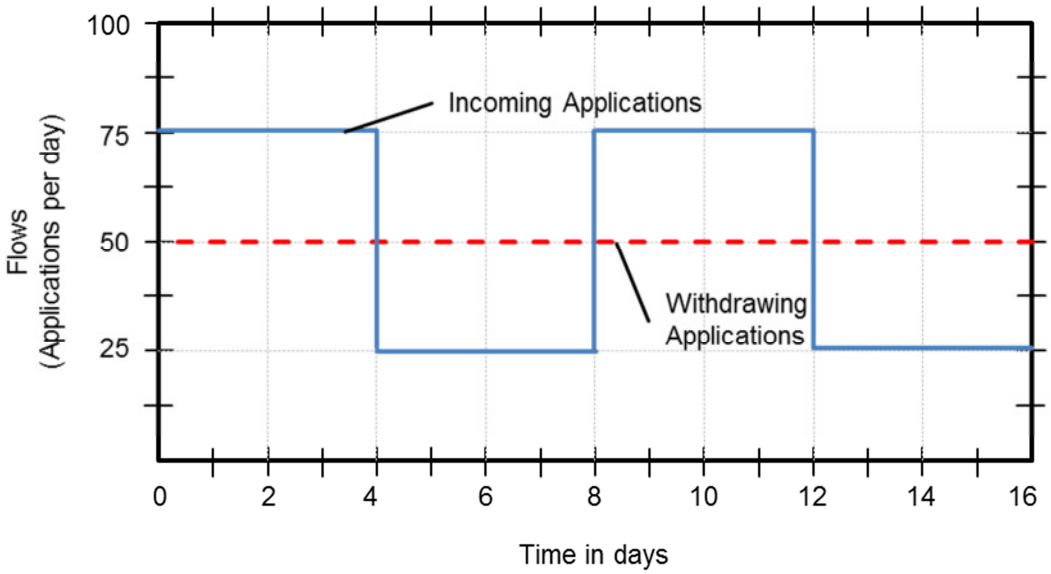


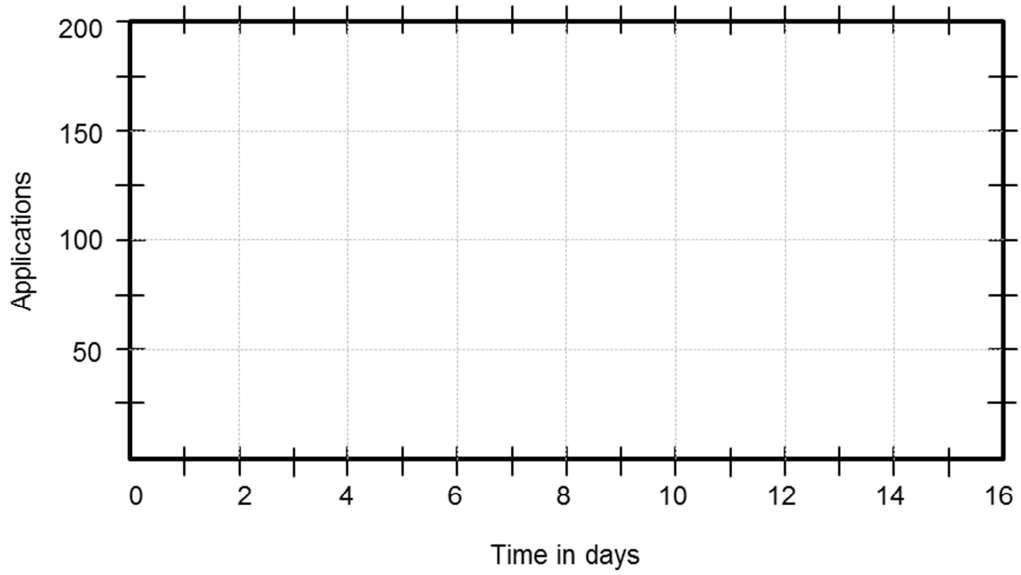
“Bathtub Dynamics” Task I (b)

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.

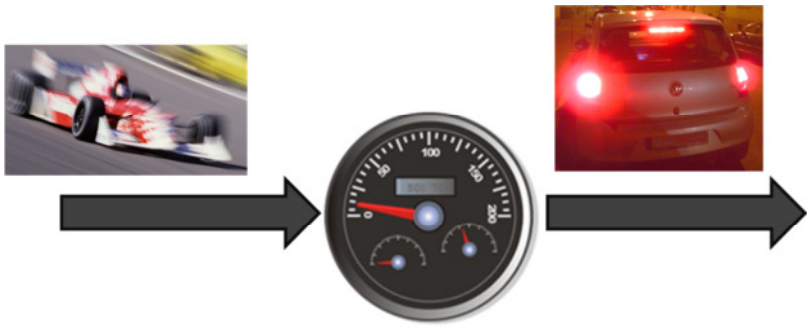
Assume the initial quantity of applications (at time zero) is 100 applications.





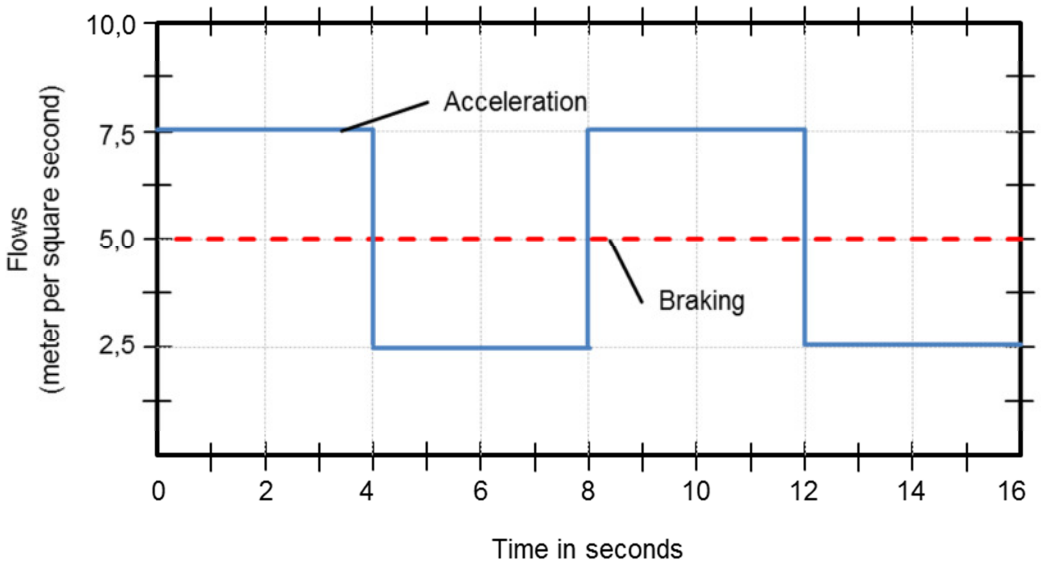
“Bathtub Dynamics” Task I (c)

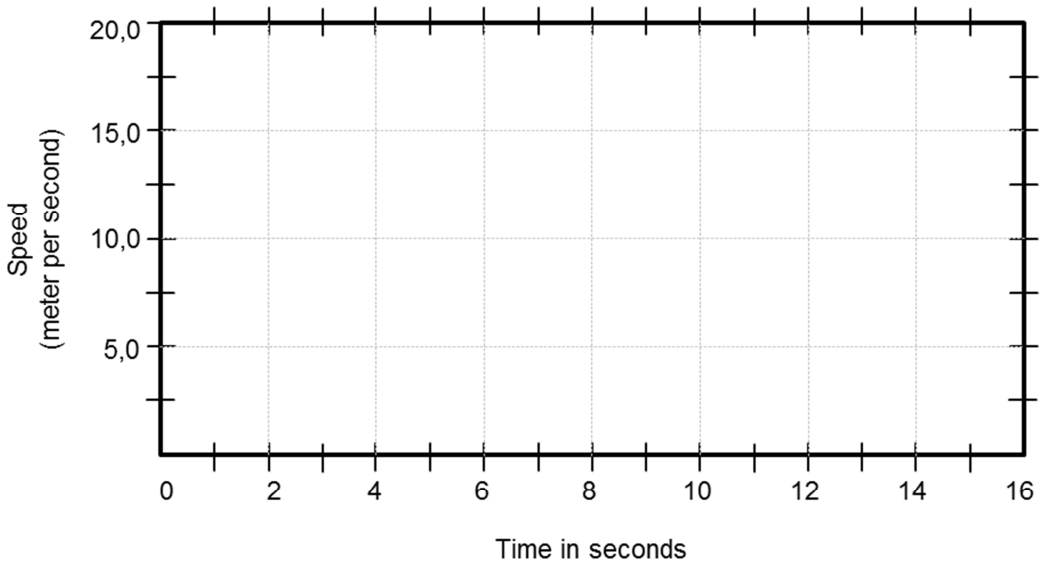
Consider the speed indicator of a car shown below. Speed increases by the car’s acceleration (shown on the left), and declines through the car braking (shown on the right):



The graph below shows the hypothetical behavior of the car accelerating and breaking. From that information, draw the behavior of the of the car’s speed on the second graph below.

Assume the initial speed (at time zero) is 10 meter per second.





“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 €.