Findings From Four Years of Bathtub Dynamics at Higher Management Education Institutions in Stuttgart

Florian Kapmeier
Lehrstuhl für Planung und Strategisches Management
Universität Stuttgart
Keplerstraße 17
70174 Stuttgart
Germany

phone: +49-(0)711-121-3465
fax: +49-(0)711-121-3191
email: florian.kapmeier@po.uni-stuttgart.de

Abstract
The 'Bathtub Dynamics' tasks, first introduced by Linda Booth Sweeney and John Sterman in 2000, have been widely used by the System Dynamics community around the world to challenge people's stock and flow thinking before being taught SD. Students at schools and universities have been taught 'Bathtub Dynamics'. The instructors' motivation was often to enlarge the sample size and hence to participate in the longitudinal analysis started by John Sterman and Linda Booth Sweeney and also to learn about their own students' systems thinking skills. We have been taking part in this ongoing research project since Fall 2000. The present paper discusses the recent results of 'Bathtub Dynamics' at the Universität Stuttgart and at the Stuttgart Institute of Management of Technology (SIMT). Overall, students' performance was poor and therefore confirms previous studies. The results contribute to the research as the two groups studied were very different regarding the demographical data and also performed differently. Also, commonalities between the Bathtub Dynamics tasks and the world-wide conducted PISA-study by the OECD are discussed.

Keywords
Bathtub dynamics, stock-flow thinking, PISA-study

Introduction

Motivated by John Sterman's presentation of his and Linda Booth Sweeney's (Booth Sweeney and Sterman 2000) study on subjects' poor understanding of stock-flow relationships at the 18th SD Conference in Bergen, Norway, we offered to join this research. In this way we were able to contribute toward increasing the sample size. Moreover, we had the opportunity to compare the initial results at MIT with the results of relatively homogeneous groups at the Universität Stuttgart and a second, more diversified group at the Stuttgart Institute of Management and Technology, SIMT. The SIMT is a joint business school of three German universities and it was recently ranked among the 30 top business schools in Europe (Quacquarelli, Saldanha, Zhang 2003).

Since 2000, roughly 120 students from the Universität Stuttgart and 31 students from the SIMT have participated in the tests. In this paper, we only present the results of the Fall 2003-2004, 2002-2003, and the Spring 2003 terms, because – as opposed to the terms before - we conducted the research using the same four 'Bathtub Dynamics' tasks at both universities. Hence, we received three comparable samples that will be discussed in this paper.

Recently, Sterman (2002) noted that research in this area has also been conducted at other schools and universities with very different groups at different ages. Yet, the overall results were rather similar and astonishing each time: performance was generally poor (Sterman 2002) – as in the present study. In the following we report on our groups' 'Bathtub Dynamics' results. We first describe the method we used to conduct the research. As the tasks are principally the same as the ones that Booth Sweeney and Sterman (2000) and Sterman (2002) describe in their articles in depth, we address the method only briefly before showing the results. Here, we compare the Universität Stuttgart 2003-2004, 2002-2003, and SIMT 2003 results to those results obtained by Booth Sweeney and Sterman (2000), Sterman (2002), Ossimitz (2002), and the preliminary results discussed by Kapmeier and Zahn (2001). Finally, we discuss the outcomes obtained.

Methods and Solutions

In this section we briefly present the method we used to conduct the research. We explored the students' understanding of stocks and flows using the four 'bathtub dynamics' tasks 'Bathtub', 'Cash Flow', 'Manufacturing' (Booth Sweeney and Sterman 2000), and 'Department Store' (Sterman 2002; Ossimitz 2002). The tasks were handed out to students of both the Universität Stuttgart and the SIMT.

In order to guarantee comparability between the research conducted so-far (Booth Sweeney and Sterman 2000; Kapmeier and Zahn 2001; Kainz and Ossimitz 2002; Ossimitz 2002; Sterman 2002; Sterman and Booth Sweeney 2002; Fisher 2003; Heinbokel and Potash 2003; Kubanek 2003; Lyneis and Lyneis 2003; Quaden and Ticotsky 2003; Zaraza 2003) and our results, the structure and the content of the original tasks were retained. The only difference was that the tasks for the students from the Universität Stuttgart were translated into German as the majority of the interviewees in this target group were German.

The first handout consisted of two challenges, the 'Bathtub' and the 'Cash Flow' tasks. In both tasks, simple patterns of in- and outflows to a single stock are given with the outflows being constant. Only the inflow patterns differ. While there were two different patterns (square wave and sawtooth pattern) presented to the MIT students for both the 'Bathtub' and the 'Cash Flow' tasks (Booth Sweeney and Sterman 2000), we decided to conduct our research with the square wave pattern only for the 'Bathtub' task and the sawtooth pattern only for the 'Cash Flow' task (see Booth Sweeney and Sterman 2000 for a detailed description).

The second challenge covers the ‘Manufacturing’ and the ‘Department Store’ cases. The ‘Manufacturing’ task (Booth Sweeney and Sterman 2000) differs from the tasks described above as it includes both a negative feedback loop and time delays.

The fourth and last task is the ‘Department Store’ (Sterman 2002). Here, a graph shows an in- and an outflow indicating the number of people entering and leaving a department store per minute over a period of 30 minutes. Students are asked to answer four questions.

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

Subjects and Procedure

As stated above, all four tasks were handed out at different times to three different groups of students with different backgrounds. Therefore, the results are presented separately in order to guarantee comparability of the three independent samples (Bortz 1999).

The first two groups of students were enrolled in the ‘System Dynamics’ course at the Universität Stuttgart. The course is offered as an elective in the field of Strategic Management during their advanced study period at master’s level. The course is also open to students from other faculties of the university. The System Dynamics courses under study were offered in the winter terms 2002-2003 and 2003-2004 and consisted of fourteen 90-minute lectures each. The last group of students was enrolled in the ‘Business Modeling’ class which is offered as an elective course in the third term of a 2-year MBA program at SIMT. This particular Business Modeling course of ten 180-minute lectures was held in the spring term of 2003.

Even though the teaching content for the groups differs slightly, the structures of both SD-classes is similar. Therefore, the tasks could be handed out to the students at comparable points in time. The first two tasks (‘Bathtub’ and ‘Cash Flow’) were distributed to both student groups on their first day of class. The second tasks (‘Manufacturing’ and ‘Department Store’) were handed out to the students 4 classes (Universität Stuttgart) or 5 classes (SIMT) later. In both classes, this was just before the students were introduced to stock-and-flow diagrams. Before answering the questions, the students were told that the purpose of the tasks was to participate in a longitudinal analysis conducted by MIT’s System Dynamics Group to gain insight into people’s understanding of stock-and-flows before they were introduced to System Dynamics. Moreover, it was stressed that the participants’ performance on the tasks would not influence the students’ grades. In addition, the students were not being paid. The interviewees had 10 minutes to work on the first task and 15 minutes for the second.

As can be seen from the Table 1, the participants were asked to fill out a background data sheet with information about their age, gender, current degree program, region of origin, first language, and whether they had played the Beer Game (Sterman 1989; Senge 1990) before. However, students at the Universität Stuttgart were not asked about their previous major because, unlike in the Anglo-Saxon university culture, in the

'classic' German university system the Vordiplom ('prediploma') is not regarded as a full degree. Therefore, students do not usually change their subject after their Vordiplom, but continue with their chosen subject. This implies that the students' highest prior degree is the high school diploma. Following the curriculum, students receive their Vordiplom after passing the required classes. Students typically pass this after the first 4 semesters of their study program.

	Universität Stuttgart 2003-2004		Universität Stuttgart 2002-2003		SIMT 2003	
Task	**Bathtub and Cash Flow**	**Manufacturing Case and Department Store**	**Bathtub and Cash Flow**	**Manufacturing Case and Department Store**	**Bathtub and Cash Flow**	**Manufacturing Case and Department Store**
Total Number of Students	43	39	32	34	22	19
Age (%)						
19-24	91	90	72	65	9	5
25-30	9	13	25	32	55	58
31-35	0	0	3	3	32	32
36 and up	0	0	0	0	0	0
Gender (%)						
Male	77	69	88	91	82	79
Female	23	31	13	8	18	21
Student Status (%)						
Prediploma / Bachelor Level	2	8	3	3	0	0
Maindiploma / Master Level	98	92	97	97	100	100
Prior Field of Study (%)						
Business/Management					23	26
Engineering					41	32
Social Science					14	16
Computer Science					5	5
Mathematics					5	5
Humanities					14	16
Highest Prior Degree (%)						
BA	0	0	0	0	36	42
BS	0	0	0	0	32	26
MA, MS, Diplom	0	0	0	0	18	16
Ph.D.	0	0	0	0	5	5
High School	100	100	100	100	0	0
BE. JD, BBA, MD, CPA	0	0	0	0	9	11
Current Field of Study (%)						
Business/Management	42	54	47	62	100	100
Engineering	53	44	38	26	0	0
Social Science	0	0	0	0	0	0
Science	0	0	0	0	0	0
Computer Science	2	3	16	12	0	0
Mathematics	0	0	0	0	0	0
Humanities	2	0	0	0	0	0
Region of Origin (%)						
North America (Aus. + NZ)	0	0	0	0	5	5
Europe	100	100	100	100	41	37
Asia and Middle East	0	0	0	0	41	42
Latin America	0	0	0	0	14	16
Africa	0	0	0	0	0	0
First Language (%)	**German**		**German**		**English**	
First language	91	87	91	94	9	5
Not first language	9	13	9	6	91	95
Beer Game Experience (%)						
Played before	2	5	0	0	0	100
Have not played	98	95	100	100	100	0

Table 1: Subject demographics - Universität Stuttgart and SIMT

As it is generally possible for students to change courses during the first weeks of the term, the total number of students participating in the tasks differs slightly. However, the proportion of male and female participants is unevenly distributed in all groups with a clear minority of female students (23% for the first and 31% for the second challenge at the Universität Stuttgart 2003-2004 group, and 13% and 9% at the 2002-2003 group, and 18% and 21% at SIMT respectively – for orientation purposes, the top-column row of the SIMT is shaded). As can be seen from the Table 1, nearly all the students at the Universität Stuttgart have passed their Vordiplom and worked on the second stage of their degree. Nearly everybody of the Universität Stuttgart 2003-2004 group (91% and 90%) and about two-thirds of the Universität Stuttgart 2002-2003 students were younger than 24 (72% and 65%) and a quarter were between 25 and 30 years old (25%

and 32%). This was a younger average age than at SIMT. Here, half of the students were between 25 and 30 (55% and 58%), and a third between 31 and 35 years old. Moreover, the majority of the interviewees from the Universität Stuttgart group were business majors (47% and 62%) – some of them were engineers (34% and 26%) and a few were computer scientists (9% and 12%). At SIMT, all participants were business majors (100%). Even though SIMT students' current field of study was homogeneous, both the students' prior field of study and their region of origin was very heterogeneous. 41% and 32% of the students' prior field of study was engineering, followed by 23% and 26% business majors, and 14% and 16% social scientists and humanities majors. Only a few were computer scientists and mathematicians (5% each). Most of the students come from Asia and the Middle East (41% and 42%) and Europe (41% and 37%). Some students come from Latin America (14% and 16%) and North America (5%). Due to this vast variety of countries of origin, English (SIMT task language) was only the first language of the minorities (9% and 5%). For the students at the Universität Stuttgart, German (Universität Stuttgart task language) was the first language for 91% and 87% (2003-2004) and 91% and 94% (2002-2003) of the students respectively. None of the Universität Stuttgart 2002-2003 participants had played the Beer Game during the course of the term. Of the 2003-2004 group, one participant and two participants had played the beer game when working on the first and the second task respectively. As this number is fairly low, no further distinctions are made between performance of interviewees with or without Beer Game experience in the following. At SIMT, nobody had played the Beer Game before working on the first task, while all of the students had played it before doing the second task.

Results

In this section we present the results of the task described above. We start with the 'Bathtub' task, continue with the 'Cash Flow' and the 'Manufacturing' task and conclude with the 'Department Store' task.

The 'Bathtub' Task, Task 1

Regarding the 'Bathtub' task, it can be seen in Table 2 that performance in all studies was poor. Taking into consideration the mean for all items, the Universität Stuttgart groups did as well (or poorly) as the students at MIT (Booth Sweeney and Sterman 2000) with roughly 83% being correct. Interestingly, this performance was much better than with the Universität Stuttgart group 2000-2001 (Kapmeier and Zahn 2001). Its average performance rated 15 percentage points lower. However, the group at SIMT did even worse with only 65% of the answers noted correctly. Ossimitz (2002) even observed an average performance of only 42% referring to a group of 154 students.

Yet, all studies reveal roughly the same pattern regarding the performance in the individual coding criteria. Just like MIT students and Stuttgart 2000-2001 students, the students of all three groups studied recently did best on stating correctly that the stock does not show any discontinuous jumps (86% Universität Stuttgart 2003-2004, 91% Universität Stuttgart 2002-2003, and 82% SIMT). The majority of the students from the groups also stated correctly the rising (86%, 88%, and 77%) and falling (86%, 84%, and 68%) of the stock at the appropriate times.

More than 80% of the Universität Stuttgart groups vs. only 64% of the SIMT students correctly drew the points in time when peaks and troughs of the stock occur. Eighty-four percent of both Universität Stuttgart student groups also sketched correctly a linear slope of the stock compared to only a little more than half of the SIMT students (59%). The results differ when it comes to the criteria that focus on calculating the net rate. Criterion 6 is SIMT students' worst item (50%). Interestingly, Universität Stuttgart students did comparatively well with 77% (2003-2004) and 78% (2002-2003), which exceeds the results reached by the 2000-2001 group by nearly 25% and those of the MIT students by 5%. Calculating the quantity added to and removed from the stock is the worst criterion in all studies. More than half of the Universität Stuttgart 2000-2001 students failed to calculate the 100 units, for example. Still, in the 2003-2004 and 2002-2003 terms, less than 30% of the Universität Stuttgart students failed to do so, whereas almost half of the SIMT students gave wrong answers.

The results of the Universität Stuttgart 2003-2004 and 2002-2003 groups do not differ much from each other. However, there is a remarkable difference between the performance of these groups and the group in 2000-2001. The better performance might be explained through the differing backgrounds of the group members. Whereas in 2000-2001 nearly 90% of the students were business majors, in the 2003-2004 group more than half and in the 2002-2003 group more than 30% were engineering and computer science majors. Interestingly, 83% of the engineering majors of the 2003-2004 group and 80% of the computer scientists majors and 74% of the engineering majors of the 2002-2003 group answered all the questions correctly.

Criterion	Performance on the Bathtub task - Universität Stuttgart 2003-2004	Performance on the Bathtub task - Universität Stuttgart 2002-2003	Performance on the Bathtub task - SIMT 2003	Performance on the Bathtub task - Universität Stuttgart 2000-2001	Performance on the Bathtub task - MIT
1. When the inflow exceeds the outflow, the stock is rising.	0.86	0.88	0.77	0.75	0.87
2. When the outflow exceeds the inflow, the stocks is falling.	0.86	0.84	0.68	0.75	0.86
3. The stock should not show any discontinuous jumps (it is continuous).	0.86	0.91	0.82	0.82	0.96
4. The peaks and troughs of the stock occur when the net flow crosses zero (i.e., at t= 4,8,12,16)	0.86	0.84	0.64	0.75	0.89
5. During each segment the net flow is constant so the stock must be rising (falling) linearly.	0.84	0.84	0.59	0.68	0.84
6. The slope of the stock during each segment is the net rate (i.e., +/-25 units/time period).	0.77	0.78	0.50	0.54	0.73
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.	0.77	0.72	0.55	0.46	0.68
Mean for all items	0.83	0.83	0.65	0.68	0.83

Table 2: Performance on the 'Bathtub' task , task 1

Sterman (2002) observed pattern matching as an often occurring error in the MIT group's results. Likewise, 12% of the Universität Stuttgart 2003-2004 students, 9% of Universität Stuttgart 2002-2003 students, and 23% of the SIMT students also matched the pattern for the stock to the inflow. This was the most typical error of both groups studied.

Interestingly, in her/his answer one participant with an engineering background differentiated between two graphs. She/he called one graph 'linear behavior' and the second 'system behavior'. She/he drew the first graph that describes the behavior of the stock (indicated as 'linear behavior') correctly. The second graph (indicated as 'system behavior'), however, reveals delays from minute 4 to 16, but no delay for the first 4

minutes. Then the 'system behavior' catches up with the 'linear behavior' again at the peak and the troughs. At these points, the student drew smooth transitions that lead to a more asymptotic slope of the 'system behavior'. Apparently the student assumed that the 'system' behaves differently from what she/he called 'linear behavior'. It seems as though she/he was indeed aware of time lags and non-linearities that exist in the real world. So she/he transferred her/his understanding of the real world to this simple task that only consists of linearities while mistakenly decoupling the direct stock-flow relations that are given in this task.

As Ossimitz (2002) observed in his study, it can also be stated here that the first criteria correlate highly (Universität Stuttgart 2003-2004: Pearson's R=1.000[i], Universität Stuttgart 2002-2003: Pearson's R=0. 878, and SIMT 2003: Pearson's R=0. 794) and significantly (Universität Stuttgart 2002-2003: p=0.000 and SIMT 2003: p=0.000) with each other. In fact, regarding the 'Bathtub' task, nearly all criteria correlate highly and significantly in the present study.

The 'Cash Flow' Task; Task 2

As in the previous surveys both study groups found the 'Cash Flow' task with the sawtooth pattern more difficult than the square wave pattern above. Overall, performance was poorer than in the previous task. According to Table 3, average performance was 52% for the Universität Stuttgart 2002-2003 students and only 42% for the Universität Stuttgart 2003-2004 students and 34% for the SIMT students. Seen generally, the mean of the Universität Stuttgart 2002-2003 students was poor. However, the performance is average when compared with the Universität Stuttgart group of 2000-2001 (45%), MIT (51%) and the Austrian groups (48%) studied by Ossimitz (2002).

	Criterion	Performance on the Cash Flow task - Universität Stuttgart 2003-2004	Performance on the Cash Flow task - Universität Stuttgart 2002-2003	Performance on the Cash Flow task - SIMT 2003	Performance on the Cash Flow task - Universität Stuttgart 2000-2001	Performance on the Cash Flow task - MIT
1.	When the inflow exceeds the outflow, the stock is rising.	0.49	0.59	0.27	0.47	0.48
2.	When the outflow exceeds the inflow, the stock is falling.	0.42	0.59	0.32	0.47	0.48
3.	The stock should not show any discontinuous jumps (it is piecewise continuous).	0.95	0.94	0.86	0.93	0.99
4.	The peaks and troughs of the stock occur when the net flow crosses zero (i.e., t = 2, 6, 10, 14).	0.49	0.63	0.32	0.35	0.39
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).	0.26	0.47	0.23	0.29	0.30
6.	The slope of the stock when the net rate is at its maximum is 50 units/period (t = 0, 8, 16).	0.16	0.28	0.23	0.32	0.52
7.	The slope of the stock when the net rate is at its minimum is -50 units/period (t = 4, 12).	0.16	0.31	0.14	0.32	0.51
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.	0.44	0.34	0.32	0.46	0.41
	Mean for all items	0.42	0.52	0.34	0.45	0.51

Table 3: Performance on the 'Cash Flow' task, task 2

Nearly half of the Universität Stuttgart 2003-2004 students and slightly less than 60% of the Universität Stuttgart students showed correctly that the stock rises (falls) when the inflow exceeds the outflow (and vice versa). Interestingly, about 70% of the SIMT students failed to do so. Ossimitz (2002) made a similar observation about his interviewees. However, 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 (49% of the Universität Stuttgart 2003-2004 group, 63% of the Universität Stuttgart 2002-2003 group, and 32% of the SIMT group). A surprisingly high number of Universität Stuttgart 2002-2003 students (47%) correctly related the net rate to the stock. This is more than twice as high as for SIMT students' performance and roughly more than 15 percentage points more than Universität Stuttgart in 2003-2004, 2000-2001, and MIT. Whereas substantially more interviewees from the MIT study correctly drew the maximum (52%) and the minimum (51%) slope of the stock, significantly fewer Universität Stuttgart 2003-2004 and Universität Stuttgart 2002-2003 students did so (16% and 16%, 28% and 31%). Whereas 86% (2003-2004) and 91% (2002-2003) of the Universität Stuttgart students recognized that there are no discontinuous jumps in the stock in the 'Bathtub' task all but 1 (95%, 2003-2004) and 2 students (94%, 2002-2003) drew the stock without any discontinuous jumps in the 'Cash Flow' task. Hence, the number approximately equals MIT students' performance (99%) and Universität Stuttgart 2000-2001 group's performance (93%). Even though this result was not as good in comparison (86%), this criterion was also SIMT students' best item. Nearly half of the Universität Stuttgart 2003-2004 but only a quarter of the Universität Stuttgart 2002-2003 and the SIMT students correctly calculated the maximum and the minimum of the stock. To do so, students simply had to calculate the area of the triangle bound by the net rate. This means that 50% and 70% of both groups failed to apply graphical integration correctly to this challenge, respectively.

As in the 'Bathtub' task, pattern matching was one of the most common errors for the groups (19% in the Universität Stuttgart 2003-2004 group and roughly 10% in the other two groups under study). Also, as in the 'Bathtub' task, many criteria correlate highly and significantly with each other, the strongest correlation being between criteria 1 (rising stock) and 2 (falling stock) (Universität Stuttgart 2003-2004: Pearson's R: 0.868 and p=0.000; Universität Stuttgart 2002-2003 and SIMT: Pearson's R=1.000), and between criteria 1, 2 and 4 (peaks and troughs) (Universität Stuttgart 2003-2004: Pearson's R=0.814 and p=0.000 and Pearson's=0.868 and p=0.000; Universität Stuttgart 2002-2003: Pearson's R=0.936 and p=0.000 for both criteria). Hence, one might assume that the subjects who follow the rule of an increasing stock when the inflow exceeds the outflow in the 'Bathtub' task will do the same in the 'Cash Flow' task and vice versa. So, there should be a correlation between criterion 1 in the 'Bathtub' and the 'Cash Flow' tasks. Yet, there is only little correlation (Universität Stuttagrt 2003-2004: Pearson's R=-0.009 and p=0.952; Universität Stuttgart 2002-2003: Pearson's R=0.265 and p=0.143). Likewise, the correlation between criterion 2 in both tasks (Universität Stuttgart 2003-2004: Pearson's R=-0.066 and p=0.672; Universität Stuttgart 2002-2003: Pearson's R=0.170 and p=0.353). In the Universität Stuttgart 2003-2004 group, criteria 6 and 7 (positive and negative slopes of the stock) highly correlate (Pearson's R=1.000) with each other.

The 'Manufacturing' Task

In the following we present the results of the 'Manufacturing' task for the Universität Stuttgart and SIMT groups consecutively. This is because (nearly) none of the Universität Stuttgart students and all of the SIMT students had played the beer game before working on the task. As the beer game experience might influence the students' answers, we measure up the results of the groups with the appropriate groups from previous studies. As illustrated by Booth Sweeney and Sterman (2000), due to having played the beer game before, students might have gained insights into the stock management system. Or, they might have just remembered the oscillating production in the game with their inventory first declining and then increasing. It cannot be ruled out that they transferred unreflectively their experience from the game to this task. Hence, the performance of the Universität Stuttgart 2003-2004 and 2002-2003 groups is compared to the performance of the Universität Stuttgart group of 2000-2001 and that of the MIT group that had not played the beer game. Likewise, we compare the SIMT (Beer Game) group's performance with the Universität Stuttgart group of 2000-2001 (BG) and the MIT (BG) group that had played the beer game.

Even though the 'Manufacturing' task is more difficult than the two tasks described above, it is still simple in content and structure. However, according to Table 4, although the average performance of the Universität Stuttgart students was poor (71% in 2003-2004 and 62% in 2002-2003), they only did slighlty worse than in the 'Bathtub' (83% and 68%) and noticeably better than in the 'Cash Flow' task (42% and 52%). The average performance level confirms the results of the Universität Stuttgart 2000-2001 sample where as many as 69% of the students drew the graphs correctly. Furthermore, the performance was much better than MIT's (33%). As at MIT, the students had not been introduced to stocks and flows at this point in time. However, they had already been introduced to qualitative causal-loop diagrams. Kapmeier and Zahn (2001) stated that the performance could be explained by the intensive training in questions that relate to manufacturing during the students' 'prediploma' courses. Related topics are discussed in at least two different courses, in 'statistics for business management majors' and 'production management of goods and services', both courses consisting of roughly 14 lectures, each of 135 minutes.

Criterion	Universität Stuttgart 2003-2004 - Beer Game: No	Universität Stuttgart 2002-2003 - Beer Game: No	Universität Stuttgart 2000-2001 - Beer Game: No	MIT - Beer Game: No	SIMT 2003 - Beer Game: Yes	Universität Stuttgart 2000-2001 - Beer Game: Yes
1. Production must start in equilibrium with orders.	0.90	0.76	1.00	0.47	0.45	1.00
2. Production must be constant prior to time 5 and indicate a lag of four weeks in the response to the step increase in orders.	0.69	0.65	0.71	0.41	0.36	0.73
3. Production must overshoot orders to replenish the inventory lost during the initial period when orders exceed production. Production should return to (or fluctuating around) the equilibrium rate of 11,000 widgets/week (to keep inventory at or fluctuating around the desired level).	0.62	0.50	0.65	0.30	0.23	0.73
4. Conservation of material: The area enclosed by production and orders during the overshoot of production (when production > orders) must equal the area enclosed by orders less production (when production < orders).	0.59	0.32	0.41	0.05	0.09	0.53
5. Does production oscillate?	0.03	0.06	n.a.	0.35	0.09	n.a.
6. Inventory must initially decline (because production < orders).	0.90	0.97	1.00	0.55	0.45	1.00
7. Inventory must recover after dropping initially.	0.85	0.85	0.76	0.43	0.36	0.93
8. Inventory must be consistent with the trajectory of production and orders.	0.44	0.26	0.29	0.06	0.18	0.53
Mean for all items	0.71	0.62	0.69	0.32	0.28	0.78

Table 4: Performance on the 'Manufacturing' task

Whereas the whole Universität Stuttgart 2000-2001 group determined correctly that production starts in equilibrium with orders, 90% of the Universität Stuttgart 2003-2004 group and only 76% of the Universität Stuttgart 2002-2003 groups did so. Around two-thirds of the Universität Stuttgart 2003-2004 and 2002-2003 groups considered the production adjustment delay. Nearly two thirds of the 2003-2004 and half of the 2002-2003 students drew a production trajectory overshooting orders. Even though this is 32 respectively 20 percentage points more than at MIT (30%), it also indicates that a third of (2003-2004) and half (2002-2003) the students failed on this issue. When it comes to the quantity of the production overshoot, nearly 60% (2003-2004) and only a third (2002-2003) of the students understood the importance of the conservation of material and drew correct trajectories. Hence, in 2002-2003 the great majority drew trajectories that did not overshoot at all or, alternatively, even overshot too much. Interestingly, when looking at the graphs for the stock, all but three (2003-2005) respectively one (2002-2003) confused student drew a declining inventory (90% and 97%). This is approximately the same percentage as in the Universität Stuttgart 2000-2001 survey but nearly twice as high as for MIT (55%). The great majority in both groups also drew an inventory trajectory that recovered (85%). However, nearly 60% of the 2003-2004 and more than 70% of the 2002-2003 students drew inventory paths that were inconsistent with the according production paths. This is roughly similar to the 2000-2001 group (29% correct) but substantially lower than for the MIT group (6% correct). This means that the majority of the groups failed to relate the two flows to the stock correctly.

Concerning the SIMT (BG) group, the results are very different from those described above. Average performance was only 31%. Here it is worth mentioning that nearly 40% of the students drew only a partially developed or no production path at all. If these participants were excluded from coding, the average performance would have been better (64%). However, all participants are taken into consideration when calculating the results, in order to receive a full picture of the students' performance.

More than half the students did not let production start in equilibrium with orders. And even more failed to include the time delay (64%). These two performances were poor compared to the Universität Stuttgart 2000-2001 group (BG) (100% and 73% correct) and MIT (BG) (57% and 46% correct). Not even one fifth of the SIMT group correctly let production overshoot orders. Except for two students, everybody failed to pay attention to the conservation of material. Furthermore, barely half of the students realized that the stock has to decline when orders are larger than production. In comparison, 77% of MIT (BG) students realized that. Roughly a third of the SIMT students let inventory recover after the initial drop. Only 18% paid attention to the consistency between the production and the inventory paths. Overall, it can be stated that the SIMT group performed more poorly than the Universität Stuttgart students.

In the 'Manufacturing' task, pattern matching and spreadsheet thinking were the two most common phenomena in both groups studied. Ten percent of the Universität Stuttgart 2003-2004 and 6% of the Universität Stuttgart 2002-2003 students drew a trajectory for inventory that just copied the production path. This indicates a complete negligence of the orders and the outflow of the system respectively. Referring to climate change, Sterman (2002) impressively describes why it is dangerous to neglect flows while following pattern matching (p. 507).

Another frequent answer pattern the students drew was what Booth Sweeney and Sterman (2000) call 'spreadsheet thinking': subjects draw short paths with steps at the end of each week for inventory. It seems as if students were confused or heavily influenced by the accounting lecture/department about the time at which decisions or observations are made – either after a certain time interval (as in spreadsheets) or continuously. Twenty-one percent of the Universität Stuttgart 2003-2004 and 9% of the Universität Stuttgart 2002-2003 students drew paths for inventory with little steps at the end of each week. These numbers are small compared to 41% of the SIMT students. In other words, nearly half of the SIMT interviewees showed this spreadsheet thinking phenomenon. Interestingly, none of the students applied spreadsheet thinking in the other tasks.

Correlation in the 'Manufacturing' task can be observed less frequently than in the previous tasks. Criterion 4 (conservation of material) is highly significant with criterion 3 (production overshooting orders) (Universität Stuttgart 2003-2004: p=0.000; Universität Stuttgart 2002-2003: p=0.000; SIMT: p=0.010). This means that subjects did consider the law of conservation of material if they had production overshooting orders before. Although there are higher and more significant correlations in each task, we only pinpoint those that are common in both tasks. Criterion 8 (consistent trajectories of inventory with production and orders) correlates highly (Universität Stuttgart 2003-2004: Pearson's R=0.0628; Universität Stuttgart: Pearson's R=0.440; SIMT: Pearson's R=0.661) and significantly (Universität Stuttgart 2003-2004: p=0.000; Universität Stuttgart 2002-2003: p=0.009; SIMT: p=0.03) with criterion 4 (conservation of material). It indicates that subjects who consider conservation of material also draw a correct inventory path that is consistent with the trajectories of production and orders.

The 'Department Store' Task

As can be seen in the Table 5, students' correct answers in the 'Department Store' task are quite similar in the three survey groups. One can also state that there are no considerable differences when comparing the performance with that of the MIT students (Sterman 2002). As suggested by Ossimitz (2002) and Sterman (2002), coding was conducted generously. This means that a student's answer was regarded as correct if the answer was within a ±1 minute range of the correct answer.

	Universität Stuttgart 2003-2004		Universität Stuttgart 2002-2003		SIMT 2003		MIT	
Criterion	**Correct Answer**	**"Cannot Be Defined"**	**Correct Answer**	**"Cannot Be Determined"**	**Correct Answer**	**"Cannot Be Determined"**	**Correct Answer**	**"Cannot Be Determined"**
1. Most people enter the store in minute 4.	0.97	0.00	0.97	0.00	1.00	0.00	0.94	0.00
2. Most people leave the store in minute 21.	0.97	0.00	1.00	0.00	0.95	0.00	0.94	0.00
3. Most people are in the store in minute 13.	0.49	0.13	0.44	0.26	0.42	0.21	0.42	0.17
4. Fewest people are in the store in minute 30.	0.36	0.23	0.26	0.38	0.26	0.21	0.30	0.28
5. Mean for all items	0.70	0.09	0.67	0.16	0.66	0.11	0.65	0.11

Table 5: Performance on the 'Department Store' task

All but one (and all respectively) student(s) of the Universität Stuttgart 2002-2003 group and everybody (all but one) in the SIMT group stated correctly when the most people enter (leave) the store. Students of the Universität Stuttgart 2003-2004 group and students at MIT answered similarly well with 97% and 94% correct. As students just had to look for the peaks of people entering and leaving on the graph, one can conclude that students can indeed read graphs properly. However, when asked for the accumulations of people entering and leaving, the participants performed significantly worse. More than half of the Universität Stuttgart (2003-2004: 49% and 2002-2003: 44% correct) and SIMT students (42% correct) failed to accumulate the number of people simply by looking for the point of interception of the two rates. Thirteen percent of the Universität Stuttgart 2003-2004 students and roughly one quarter of the Universität Stuttgart 2002-2003 students and slightly fewer SIMT students stated that the answer cannot be determined. Nevertheless, some participants, although stating that the answer cannot be determined, noted comments that demonstrate that they were on the right track. One student mentioned, for example, that the minute in which most people are in the store could be determined by calculating the "the sum of entering – the sum leaving. But we cannot calculate it within the time given." Another participant had the right idea, but stated that "we cannot calculate the difference between entering and leaving offhand." So, both had the right intuition, but they did not use the graphical approach to solve the challenge. One student made a more personal statement, giving the answer to question 3 that most people are in the store "… when I am in a rush." Interestingly, nearly one third of the Universität Stuttgart 2002-2003 students and slightly fewer Universität Stuttgart 2003-2004 (26%) and SIMT students (27%) noted that the most people were in the store in minute 8. This is the minute when the difference between the two flows is at its maximum. Moreover, it is also the minute when the least people are leaving. This means that these participants did not manage to differentiate between an accumulation and the net rate.

The results get worse when asked for the fewest people in the store. Here, only roughly a third (Universität Stuttgart 2003-2004) and a quarter (Universität Stuttgart 2002-2003 and SIMT) of the groups answered correctly. As 23% (Universität Stuttgart 2003-2004), 38% (Universität Stuttgart 2002-2003), and 21% (SIMT) of the students indicated that

the answer “cannot be determined”, around a third of the Universität Stuttgart students and more than half of the SIMT students gave incorrect answers. Again, 10% of the Universität Stuttgart 2003-2004 and nearly one third of Universität Stuttgart 2002-2003 students answered that the fewest people are in the store in minute 17. As above, this is the moment when the difference between the two flows is at its maximum after the crossing of the two flows, but here with the least people entering. Interestingly, it is not only the same percentage but also exactly the same subjects (except for 2) who indicated minute 8 for question 3 in both Universität Stuttgart groups. Consequently, this confirms that the subjects who failed to differentiate between a net rate and an accumulation in question 3 were indeed confused as they continued to mix them up in question 4.

Regarding correlation in the ‘Department Store’ task it can be concluded that those participants who correctly determined the time when most people are in the store also determined correctly the minute when the least people are in the store (Universität Stuttgart 2003-2004: Pearson’s R=0.661 with p=0.000; Universität Stuttgart 2002-2003: Pearson’s R=0.675 with p=0.000; SIMT: Pearson’s R=0.701 with p=0.001). Unfortunately, no Pearson’s R statement can be made regarding the correlation between criterion 1 and 2 in any of the groups as there are no variations in the answers.

Impact of Subject Demographics

Booth Sweeney and Sterman (2000) observed an impact of subject demographics like prior academic background and region of origin as highly significant and the prior degree as significant. Due to the relatively small number of participants in the present study, it is not reasonable to look for significance. However, it is possible to observe some impact of subject demographics on performance. As can be seen from the Table 1, we only consider the criteria that show variety in the demographics (see Table 6). Hence, for the Universität Stuttgart groups we only concentrate on age and current program and for the SIMT group we consider age, previous major, highest previous degree, and the region of origin. Due to the uneven distribution of males and females in this study, we cannot make any suggestions concerning the gender effect noticed in Booth Sweeney’s and Sterman’s (2000), Ossimitz’s (2002), or Kainz’ and Ossimitz’ (2002) studies. Yet, interestingly, unlike in the studies of Booth Sweeney and Sterman (2000) and Lyneis and Lyneis (2003), age does have an impact on the performance in some tasks in the present groups. It can be concluded that subjects perform better the older they are (Cramer’s $V^{ii} > 0.5$ in the ‘Department Store’ task in the Universität Stuttgart 2003-2004 group and in all other cases except for the ‘Manufacturing Case’ in the Universität Stuttgart 2002-2003 group). A reason for this observation might be the structure of the sample group (see Table 1). There are more older subjects in the SIMT group than in the Universität Stuttgart group.

Uni Stuttgart 2003-2004	Bathtub		Cash Flow		Manufacturing		Department Store	
Variable	Cramer's V	*p*	Cramer's V	*p*	Cramer's V	*p*	Cramer's V	*p*
Age	0.363	0.713	0.462	0.072	0.446	0.154	*0.564*	0.002
Current Program	0.308	0.906	0.479	0.1	0.473	0.149	*0.771*	0.000

Uni Stuttgart 2002-2003	Bathtub		Cash Flow		Manufacturing		Department Store	
Variable	Cramer's V	*p*	Cramer's V	*p*	Cramer's V	*p*	Cramer's V	*p*
Age	*0.6*	0.005	*0.558*	0.027	0.396	0.869	*0.514*	0.172
Current Program	0.394	0.621	0.464	0.468	*0.564*	0.42	0.228	0.74

SIMT 2003	Bathtub		Cash Flow		Manufacturing		Department Store	
Variable	Cramer's V	*p*	Cramer's V	*p*	Cramer's V	*p*	Cramer's V	*p*
Age	*0.659*	0.652	*0.697*	0.433	*0.766*	0.84	*0.602*	0.719
Previous Major	*0.526*	0.21	0.45	0.843	*0.587*	0.138	0.407	0.854
Previous Degree	0.491	0.383	0.49	0.629	0.49	0.572	0.343	0.876
Region of Origin	*0.635*	0.032	*0.501*	0.555	*0.806*	0.01	0.419	0.348

Table 6: Impact of subject demographics on performance (items with high correlation (Cramer's V > 0.5) in *italics*)

As the Universität Stuttgart groups consisted of participants with different current majors, we also tested this criterion. Although we do not find any indication for correlation at most tasks, there is a correlation between the current program and the performance in the 'Department Store' task (Cramer's V=0.771). As can be seen from the Table 7, business majors performed best (mean=3.14, standard deviation=0.96). In the SIMT group the previous major had a significant influence. Its impact is observable on the performance in the 'Bathtub' (Cramer's V=0.526) and the 'Manufacturing' (0.587) tasks. As can be seen from the Table 7, former engineers performed best with a mean of 5.67 (standard deviation=2) correct answers out of 7 in the 'Bathtub' task, and the computer scientist and the mathematician performed best (mean=5.00) in the 'Manufacturing' task. Interestingly, the two social scientists performed diversely. The standard deviation in the 'Bathtub' and the 'Manufacturing' task equals to 4.95 (mean=3.50) and 4.95 (5.00) respectively.

Performance depended on the participants' region of origin in the 'Bathtub' (Cramer's V=0.635), the 'Cash Flow' (Cramer's V=0.501), and the 'Manufacturing' task (Cramer's V=0.806). Interestingly, the subjects originating from Europe performed best on average in the first three tasks. Subjects from North and Latin America achieved the highest average score on the 'Department Store' task.

Uni Stuttgart 2003-2004

Bathtub

Current Program	μ	N	σ
Business / Management	5.17	18.00	2.92
Engineering	6.22	23.00	2.04
Computer Science	**7.00**	1.00	.
Philosophy	**7.00**	1.00	.
Total	5.81	43.00	2.44

Gender	μ	N	σ
Male	**6.36**	33.00	1.75
Female	4.00	10.00	3.50
Total	5.81	43.00	2.44

Uni Stuttgart 2002-2003

Bathtub

Current Program	μ	N	σ
Business / Management	5.80	15.00	2.51
Engineering	**5.92**	12.00	2.11
Computer Science	5.60	5.00	2.61
Total	5.81	32.00	2.31

Cash Flow

Current Program	μ	N	σ
Business / Management	2.94	18.00	2.31
Engineering	3.65	23.00	2.76
Computer Science	2.00	1.00	.
Philoyophy	**6.00**	1.00	.
Total	3.37	43.00	2.55

Gender	μ	N	σ
Male	**3.85**	33.00	2.56
Female	1.80	10.00	1.81
Total	3.37	43.00	2.55

Cash Flow

Current Program	μ	N	σ
Business / Management	4.20	15.00	2.62
Engineering	**4.83**	12.00	2.95
Computer Science	2.40	5.00	3.21
Total	4.16	32.00	2.86

Manufacturing

Current Program	μ	N	σ
Business / Management	**5.33**	21.00	1.80
Engineering	4.94	16.00	2.41
Computer Science	0.00	1.00	.
Total	5.03	38.00	2.20

Gender	μ	N	σ
Male	4.89	27.00	2.22
Female	**5.36**	11.00	2.20
Total	5.03	38.00	2.20

Manufacturing

Current Program	μ	N	σ
Business / Management	4.00	17.00	2.00
Engineering	**4.83**	12.00	1.75
Computer Science	4.20	5.00	2.49
Total	4.32	34.00	1.97

Department Store

Current Program	μ	N	σ
Business / Management	**3.14**	21.00	0.96
Engineering	2.56	16.00	0.73
Computer Science	0.00	1.00	.
Total	2.82	38.00	1.01

Gender	μ	N	σ
Male	**3.15**	27.00	0.86
Female	2.00	11.00	0.89
Total	**2.82**	38.00	1.01

Department Store

Current Program	μ	N	σ
Business / Management	2.71	17.00	0.85
Engineering	2.58	12.00	1.00
Computer Science	**2.80**	5.00	1.10
Total	2.68	34.00	0.91

SIMT 2003

Bathtub

Previous major	μ	N	σ
Business / Management	4.50	6.00	2.81
Engineering	**5.67**	9.00	2.00
Social Science	3.50	2.00	4.95
Computer Science	5.00	1.00	.
Mathematics	0.00	1.00	.
Humanities	3.33	3.00	3.21
Total	4.55	22.00	2.72

Highest previous degree	μ	N	σ
BA	4.75	8.00	3.15
BS	4.86	7.00	2.48
MA, MS, Diplom	3.25	4.00	3.30
PhD	**5.00**	1.00	.
BE, JD, other	**5.00**	2.00	2.83
Total	4.55	22.00	2.72

Region of origin	μ	N	σ
North America (incl. Aus, NZ)	2.00	1.00	.
Europe	**5.56**	9.00	2.30
Asia & Middle East	4.00	9.00	3.16
Latin America	4.00	3.00	2.65
Total	4.55	22.00	2.72

Cash Flow

Previous major	μ	N	σ
Business / Management	2.67	6.00	2.94
Engineering	3.22	9.00	2.44
Social Science	1.50	2.00	0.71
Computer Science	**4.00**	1.00	.
Mathematics	1.00	1.00	.
Humanities	2.00	3.00	1.00
Total	2.68	22.00	2.23

Highest previous degree	μ	N	σ
BA	2.50	8.00	2.56
BS	1.86	7.00	1.07
MA, MS, Diplom	**4.00**	4.00	3.16
PhD	**4.00**	1.00	.
BE, JD, other	3.00	2.00	2.83
Total	2.68	22.00	2.23

Region of origin	μ	N	σ
North America (incl. Aus, NZ)	1.00	1.00	.
Europe	**4.00**	9.00	2.74
Asia & Middle East	2.11	9.00	1.36
Latin America	1.00	3.00	0.00
Total	2.68	22.00	2.23

Manufacturing

Previous major	μ	N	σ
Business / Management	1.40	5.00	1.95
Engineering	3.14	7.00	2.48
Social Science	3.50	2.00	4.95
Computer Science	**5.00**	1.00	.
Mathematics	**5.00**	1.00	.
Humanities	0.33	3.00	0.58
Total	2.47	19.00	2.52

Highest previous degree	μ	N	σ
BA	1.50	8.00	2.62
BS	2.40	5.00	2.30
MA, MS, Diplom	**4.00**	3.00	3.61
PhD	**4.00**	1.00	.
BE, JD, other	3.50	2.00	0.71
Total	2.47	19.00	2.52

Region of origin	μ	N	σ
North America (incl. Aus, NZ)	1.00	1.00	.
Europe	**5.00**	7.00	1.41
Asia & Middle East	1.00	8.00	1.93
Latin America	1.00	3.00	1.73
Total	2.47	19.00	2.52

Department Store

Previous major	μ	N	σ
Business / Management	**3.00**	5.00	1.00
Engineering	2.71	7.00	1.25
Social Science	2.50	2.00	0.71
Computer Science	2.00	1.00	.
Mathematics	2.00	1.00	.
Humanities	2.33	3.00	0.58
Total	2.63	19.00	0.96

Highest previous degree	μ	N	σ
BA	2.38	8.00	0.92
BS	2.60	5.00	0.89
MA, MS, Diplom	2.67	3.00	1.15
PhD	**4.00**	1.00	.
BE, JD, other	3.00	2.00	1.41
Total	2.63	19.00	0.96

Region of origin	μ	N	σ
North America (incl. Aus, NZ)	**3.00**	1.00	.
Europe	2.57	7.00	1.13
Asia & Middle East	2.50	8.00	0.93
Latin America	**3.00**	3.00	1.00
Total	2.63	19.00	0.96

Table 7: Mean performance separated according to different demographic data (items with highest mean performance in **bold**)

Discussion

The results of the two 'Bathtub Dynamics' challenges, each consisting of two tasks, conducted at the Universität Stuttgart and SIMT mostly correspond with the results obtained at MIT. Therefore, our explanations for the results mainly coincide with those of Booth Sweeney and Sterman (2000) and Sterman (2002).

However, we would like to go into detail in regard to specific topics. As at MIT, we presented highly educated students at the Universität Stuttgart and at SIMT with the 'Bathtub Dynamics' challenges. The goal was to test the subjects' understanding of stock-and-flow structures before they were introduced to System Dynamics and quantitative SD.

Except for a few participants, all subjects from the Universität Stuttgart groups and the SIMT group were at Master's level. Half of the Universität Stuttgart 2003-2004 group and the majority of the Universität Stuttgart 2002-2003 group studied business administration, some studied engineering and a few computer science. During high school all subjects were educated in Europe, most of them in Germany. However, when interpreting the results one has to keep in mind that over the last 10 years it has been observed that students who attend the elective SD course at the Universität Stuttgart are typically more ambitious than the students who choose to attend one of the other alternative elective courses. At SIMT, all students were enrolled in an MBA program. Here, almost 60% come from Asia, Latin America, and North America. Generally speaking, the three groups' overall performance on the challenges was poor.

The results from the Universität Stuttgart groups from 2003-2004, 2002-2003, and 2000-2001 do not differ very much. It can be said that the preliminary results from 2000-2001 (Kapmeier and Zahn 2001) have been confirmed to a certain extent. However, the mistakes made are not calculation errors. Subjects do profoundly violate fundamental stock-and-flow relationships. First, subjects violated the conservation of matter. Nearly half of the Universität Stuttgart 2003-2004 and two-thirds of both the Universität Stuttgart 2002-2003 and SIMT groups drew trajectories for the stock that were inconsistent with the net rate. Second, pattern matching was one of the main errors. In the 'Bathtub' task, for example, where the inflow is discontinuous, 18% of the SIMT subjects copied the inflow to draw the stock. In the 'Cash Flow' task, where the inflow is continuous, only 14% of the SIMT group drew a discontinuous stock.

However, even though overall performance on the 'Manufacturing' task was poor, Universität Stuttgart subjects did perform relatively well on it (71% of the Universität Stuttgart 2003-2004 and 63% of the Universität Stuttgart 2002-2003 group). This finding is about that for the Universität Stuttgart 2000-2001 group (69% without Beer Game and even 78% with Beer Game) but it is still better than the MIT group (33% and 46%) and the SIMT (BG) group (28%). Hence, Kapmeier's and Zahn's (2001) hypothesis that this result could be traced back to extensive training in production matters during the students' 'prediploma' can be supported. This could also be an explanation for the relatively good results of the Universität Stuttgart students in comparison to the SIMT students in the 'Bathtub', 'Cash Flow', and 'Manufacturing' tasks. However, this observation will be investigated further in future courses.

For the ‘Department Store’ task it can be stated that the results of both groups lie within the range of results obtained by Sterman (2002) and Lyneis and Lyneis (2003), even though the students from the latter group had already been introduced to stock-and-flow structures when working on this task (Lyneis and Lyneis 2003). From the findings it can be stated that subjects do understand how to read graphs but fail to accumulate flows.

Indeed, it could be assumed that the students were not motivated enough and hence did not put much effort into working on the tasks as they did not receive any incentives, i.e., grades or money (Booth Sweeney and Sterman 2000). However, as the knowledge requirements to work on the challenges are comparatively straightforward, they therefore should be routine. Also, as the students in the original tasks were not paid either, the findings of the surveys are comparable.

On the one hand, Booth Sweeney and Sterman (2000) and Sterman and Booth Sweeney (2002) point towards the naturalistic decision-making perspective that the results could be poor due to the unfamiliarity of the situations in the tasks. However, as the authors indicate, filling a bathtub, checking a bank account and observing how many people enter and leave a store are situations that not only students are confronted with in their everyday lives. Considering the ‘Manufacturing’ task, one could claim that the situation described is a content of their studies and consequently they should be familiar with it.

On the other hand, the authors (Booth Sweeney and Sterman 2002) point out that according to the evolutionary perspective, people do not need to understand the relationships between stocks and flows as nature always accumulates stocks properly. This might indeed be correct, but people should have a feeling for flows and their related accumulations – not only for the production decisions in the company system but also for understanding the impacts of decisions on the development of the ecological system (see Booth Sweeney and Sterman 2002 or Wackernagel and Rees 1995, for example).

The results of the survey indicate that the tasks were too extensive for the time given, as one student remarked. Hence, this matter should be observed in future tasks.

As stated above, both sample groups in this study are too small to look for significant influence of demographic data on the performance. Yet, we found some indications of correlation between the two. Age, prior field of study, the highest previous degree, and the region of origin correlate highly with performance. However, these findings are still preliminary.

It could, for example, be a subject for future research to evaluate the correlation between performance in the ‘Bathtub Dynamics’ tasks and the region of origin in respect to the findings of the “Programme for International Student Assessment” (PISA) study. The PISA study is part of the “Indicator Programme” of the Organisation for Economic Co-operation and Development (OECD). Its goal is to obtain comparable data on the countries’ education systems, asking high school students not for factual knowledge but for base competencies that are seen as crucial for people to take part in social, economic, and political life. Three different competence areas stand in the focus of the study: reading, mathematical, and scientific literacy. All areas are tested in three

study cycles being conducted in 2000, 2003, and 2006. In each assessment, there is a special focus on one of the three domains. In the first study cycle, conducted in 2000, the focus lay on reading literacy. Roughly 200,000 15-year-old high school students in more than 30 countries participated in the first study cycle (OECD 2000; Stanat et al. 2002).

Comparing the PISA study with the 'Bathtub Dynamics', at least three analogies can be identified. The first analogy refers to a more superordinate goal. There is a similarity in the information that the OECD and System Dynamicists seek for when conducting their analyses. The PISA study measures how well young adults are prepared to meet the challenges of today's knowledge societies at the end of compulsory schooling. Looking at the PISA study definitions of the three fields of interest gives us a hint as to the superordinate goal of the study. Reading literacy, for example, measures the understanding of "written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society" (OECD 2003, p. 22). Mathematical literacy is defined as the ability to "make well-founded judgments about the role that mathematics plays, as needed for an individual's current and future life, occupational life, social life with peers and relatives, and life as a constructive, concerned, and reflective citizen" (OECD 2003, p. 82). Scientific literacy is seen as the capacity "to understand and help make decisions about the natural world and the changes made to it through human activity" (OECD 2003, p. 102). To summarize, the PISA study challenges the constructive, concerned, and reflective citizen who learns to participate in society and who makes decisions about the reality that she/he is part of. It is comparable to what System Dynamicists are searching for. System Dynamics supports learning for an insightful understanding of complex systems that we live in to solve real world problems by making decisions that do not suffer from policy-resistance (Sterman 2000). Hence, there is an analogy between the PISA study's challenge and what Systems Thinkers and System Dynamicists proclaim.

Second, the skills being tested in both tests are comparable. The PISA study tests 15-year olds' critical and reflective decision-making. The assessment focuses on the students' ability to apply their knowledge and skills to meet real-life challenges, rather than on the extent to which they have mastered a specific school curriculum (OECD 2003). Among others, students are asked to transfer the knowledge learned at school to different contexts of application, i.e., using scientific concepts like physical changes, forces and movement, or human biology, and apply them to life and health, science in the environment, or science in technology. Looking at the 'Bathtub Dynamics' tasks, we also test the interviewees' ability to apply their knowledge, and specifically the systems thinking skills which have been learned at school and, if applicable, at further education institutions, to very similar fields. Hence, both tests have in common the examination of the current understanding of our everyday life.

Third, keeping in mind the areas dealt with in the 'Bathtub Dynamics' tasks, there are similarities concerning the content of both kinds of tests. The areas dealt with in the PISA study, for example, include the water level of the disappearing and reappearing Lake Chad in Africa, the human immune system and the flu, the speed of a racing car, infections, antibiotic resistance, or the ozone layer. The students are required to draw graphs, compile tables, or read and interpret texts, for example. Interestingly, the areas

under discussion often involve stocks and flows. However, the students are not asked about these directly, but more subtly. For instance, students are given a graph with information on the actual speed of a racing car in km/h over a racing track of 3 km. In stock-and-flow terms, the speed can be interpreted as a flow that fills up a stock called mileage. When the subjects are asked what the lowest speed is, they are simply being asked for the lowest point of the graph or the flow respectively. Students are also asked whether the speed of the car increases, decreases or stays constant, referring to the slope of the flow. Furthermore, in the example of Lake Chad, students are asked about a stock. Here, a graph is given showing the water level. Students are then asked to estimate the water level – the stock - at a certain point in time (OECD 2003). Summing up, there is a strong overlap of both topics and basic stock-and-flow thinking in the questions in the PISA study and the 'Bathtub Dynamics' tasks.

According to the PISA findings, US American or German high school students are ranked much lower than their counterparts in Finland or Japan. Hence, it would be interesting to investigate whether subjects who perform poorly in the 'Bathtub Dynamics' tasks originate from countries that are ranked lower in the PISA study than the countries of high-performers. There is a strong analogy between the two studies. Therefore, our proposition would be – and that would be a subject for future research – that there is a correlation between the results of the studies.

It would also be a subject for future research to do the 'Bathtub Dynamics' tasks with participants who are not necessarily business majors or engineers but students of humanities, arts, and social sciences or high school students. Sterman (2002) reports that in the meantime 'Bathtub Dynamics' have been studied in high schools in the US and Canada. The instructors presented their results with high-school students at the 2003 SD Conference in NYC (Fisher 2003; Heinbokel and Potash 2003; Kubanek 2003; Lyneis and Lyneis 2003; Quaden and Ticotsky 2003; Zaraza 2003). Generally speaking, the results resemble both ours and those of MIT.

To sum up, in order to conduct the research suggested above a more sophisticated background information sheet should be handed out. It would capture questions that give relevant hints about the participants' educational and cultural background. First, the sheet would need to ask questions relating to primary school education. They should, for example, be linked to the findings from the PISA study in order to challenge the hypothesis that students who went to high school in Japan or Korea – the countries whose students performed best in the mathematical and scientific literacy tests - also perform better in the 'Bathtub Dynamics' tasks than students who went to high school in the US or Germany – countries where student performance is only average in the PISA study. If the participants have already passed their high school exam they could also be asked whether mathematics was one of their majors during their final year in high school. Second, other questions could relate to their secondary school education, depending on the interviewees' status of school education. The questions should ask for estimates of how many hours, i.e. in semester periods per week, they have been trained in mathematics in general, and specifically in calculus and graphical analysis so far.

In this way the international comparability and the analysis of 'Bathtub Dynamics' tasks would be further increased. It would enable us to put the findings into perspective with

findings of internationally renowned studies such as the PISA study. It would also enable us to look for the roots of the relatively poor results in the 'Bathtub Dynamics' tasks. We may receive more profound hints about the connection between training in calculus and performance in the 'Bathtub Dynamics' tasks. It may also give us further indication as to whether training in System Dynamics or just more training in mathematics may improve people's stock-and-flow thinking.

Interestingly, Kainz and Ossimitz (2002) found that even a crash course in SD can enhance people's understanding of stock and flows. Fisher's (2003) and Zaraza's (2003) findings with high school students and also Lyneis' and Lyneis' (2003) results with undergraduate students support this hypothesis. Fisher's students with a System Dynamics modeling background, for example, performed extremely well in the Bathtub Tasks 1 and 2 compared to those without this particular prior course work. This indication should strengthen our belief to continue to teach SD to make people more sensitive towards our social, economic, and ecological environment so that people learn to design suitable policies. One of our students gave us a first hint towards this objective when he said: "Since I took this SD course, I have been seeing the world with different eyes."

Acknowledgements
We thank Meike Tilebein, Jeroen Struben, Rachel Teear, and Matthias Fifka and an anonymous reviewer for their helpful suggestions. We thank Erich Zahn for allowing us to use some time of his classes for running these tasks, Katja Neller for her helpful advice in statistics and Stefan Grösser for supporting us with reviewing the coding. Finally, we thank the System Dynamics students of the Universität Stuttgart and the Stuttgart Institute of Management and Technology for participation in this survey.

References

Booth Sweeney L, Sterman JD. 2000. Bathtub dynamics: initial results of a systems thinking inventory. *System Dynamics Review* **16**(4): 249-294

Bortz J. 1999. Statistik für Sozialwissenschaftler. Vol. 5. Springer: Berlin.

Fischer, DM. 2003. Student performance on the bathtub and cash flow problems. *Proceedings of the 2003 International System Dynamics Conference*. System Dynamcis Society: Albany, NY.

Heinbokel J, Potash J. 2003. Bathtub dynamics at Vermont commons schools. *Proceedings of the 2003 International System Dynamics Conference*. System Dynamcis Society: Albany, NY.

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

Kapmeier F, Zahn EOK. 2001. Bathtub dynamcis: results of a systems thinking inventory at the Universität Stuttgart, Germany [online]. *Stuttgart: Universität Stuttgart, Lehrstuhl für Planung*, 2001 [cited July 1 2003]. Available from World Wide Web: (http://www.bwi.uni-stuttgart.de/fileadmin/abt4/download/rest/bathtub_dynamics_analysis_universitaet_stuttgart.pdf)

Kubanek G. 2003. Bathtub dynamics – Ottawa. *Proceedings of the 2003 International System Dynamics Conference*. System Dynamcis Society: Albany, NY.

Lyneis J, Lyneis D. 2003. Bathtub dynamics at WPI. *Proceedings of the 2003 International System Dynamics Conference*. System Dynamcis Society: Albany, NY.

Organisation for Economic Co-operation and Development (OECD) 2000. *Measuring Student Knowledge and Skills: The PISA 2000 Assessment for Reading, mathematical and scientific literacy*. OECD: Paris.

Organisation for Economic Co-operation and Development (OECD) 2003. Programme for international student assessment. Sample tasks from the PISA 2000 assessment of reading, mathematical and scientific literacy [online]. *OECD* [cited November 11 2003]. Available from World Wide Web: (http://www.PISA.oecd.org/Docs/Download/PISA-Sampleitems_L.pdf)

Ossimitz G. 2002. Stock-flow thinking and reading stock-flow-related graphs: an empirical investigation in dynamics thinking abilities. *Proceedings of the 2002 International System Dynamics Conference*. System Dynamcis Society: Albany, NY.

Quaden R, Ticotsky A. 2003. Bathtub dynamics at Carlisle Public Schools. *Proceedings of the 2003 International System Dynamics Conference*. System Dynamcis Society: Albany, NY.

Quacquarelli N, Saldanha M, Zhang Y, ed. 2003. *International Recruiters Survey 2003. Global 100 Top Business Schools*, n.p.: London

Senge P. 1990. *The Firth Discipline: The Art and Practice of the Learning Organization*. Doubleday: New York, NY.

Stanat P, Artelt C, Baumert J, Klieme E, Neubrand M, Prenzel M, Schiefele U, Schneider W, Schümer G, Tillmann K-J, Weiß M. (Eds.) *PISA 2000: Die Studie im Überblick. Grundlagen, Methoden und Ergebnisse*. Max-Planck-Institut für Bildungsforschung: Berlin.

Sterman JD. 1989. Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment. *Management Science* **35**(3): 321-339

Sterman JD. 2000. *Business Dynamics. Systems Thinking for a Complex World.* Irwin McGraw-Hill: Boston.

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

Sterman JD, Booth Sweeney L. 2002. Cloudy skies: assessing public understanding of global warming. *System Dynamics Review* **18**(2): 207-240

Wackernagel M, Rees WE. 1995. *Our Ecological Footprint: Reducing Human Impact on the Earth.* New Society Publishers: Gabriola Island, BC

Zaraza, R. 2003. Bathtub dynamics in Portland at SyMFEST. *Proceedings of the 2003 International System Dynamics Conference*. System Dynamcis Society: Albany, NY.

[i] Pearson's R is a correlation coefficient that expresses the degree of linear relationship between two variables measured from the same individual. Pearson's R values can range

between -1 to +1. A Pearson's R of +1 signifies a perfect positive relationship, while -1 shows a perfect negative relationship. The smallest correlation is zero (Bortz 1999).

[ii] Cramer's V is a transformation of the chi-squared statistic into the zero to one interval and is useful for comparing the relative intensity of association between marker pairs. Cramer's V values can range from 0 (no association) to 1 (highest possible association) (Bortz 1999).

Back to the Top