Go Back

Stock-Flow-Thinking and Reading stock-flow-related Graphs: An Empirical Investigation in Dynamic Thinking Abilities

Günther Ossimitz
University of Klagenfurt, Austria
Department of Mathematics Education
Universitätsstr. 65
A-9020 Klagenfurt Austria/Europe
phone: +43-463-2700-3132
fax: +43-463-2700-3199
ossimitz@bigfoot.com
http://guenther.ossimitz.at

Preliminary Version January 2001

Prepared for the 2001 System Dynamics Conference, Atlanta, Georgia

Stock-Flow-Thinking and Reading stock-flow-related Graphs: An Empirical Investigation in Dynamic Thinking Abilities

Günther Ossimitz
University of Klagenfurt, Austria
Department of Mathematics Education
ossimitz@bigfoot.com

ABSTRACT

This paper reports on an empirical study of the ability of university students to discern between stocks and flows, and as such is a sequel to the "Bathtub Dynamics" – study of Sweeney and Sterman (Sweeney/Sterman 2000). The 154 subjects were given six different tasks. Some of the tasks required the ability to read and interpret time-graphs, others did not. Two of the tasks were taken directly from the Sweeney/Sterman (2000) study.

The findings of this study were alarming. On the first task, which dealt with the difference between federal budgetary deficit (a net-flow) and public debt (a stock), the mean performance of the candidates was approximately on the same level as if they had flipped a coin for each answer. Generally the study showed severe deficits in the ability to discern between stocks and flows. These deficits did not depend upon whether the tasks required the candidates to read or to draw graphs or not. Graphs showing the development of flows over time were interpreted by many subjects as if these were graphs of stocks.

In almost all tasks the females scored significantly poorer than the males. At the present stage of investigation I have no reasonable explanation for this massive gender bias. Further research might help to reveal something of this mystery.

A number of extreme correlations between different criteria concerning the performance of the bathtub tasks (which have been taken from Sweeney/Sterman) suggest that there might exist a core ability which triggers all other stock-flow-thinking capabilities: this is the ability to grasp that a positive net-flow results in an increase of the corresponding stock.

The study revealed serious deficits in stock-flow-thinking, which urgently need an educational "therapy". It is a major challenge for systems-oriented education to design and evaluate appropriate courses.

1 Stock-Flow-Thinking

Discerning between stocks and flows is not only important for System Dynamics (SD) modelling (see Forrester 1961, Richmond 1993, Sterman 2000), but can be considered a genuine component of Systems Thinking (ST) abilities, too[1]. This is argued in some detail by Sweeney/Sterman (2000) or Ossimitz (2000 pp 52ff), who identifies four basic ST dimensions:

- *Thinking in Interrelated Structures* (*"vernetztes Denken"*).
- *Dynamic Thinking*, which means a thinking which is not restricted to grasping just snapshots of a situation, but takes into account evolution over time.
- *Thinking in Models*, which means that any systems thinker should be aware that he or she is always dealing with a model of a complex situation, which is usually massively simplified compared with the "actual" situation.
- *Systemic Action*, which means the practical ability of steering systems.

I will refer to the ability to discern between stocks and flows as "Stock-Flow-Thinking". Stock-Flow-Thinking is an important aspect of both the "Dynamic-Thinking" and "Thinking in Models" dimension of ST.

Stock-Flow-Thinking is a natural ability for any SD modeler and actually the ability to make models by distinguishing between stocks and correlating flows is in the very core of the SD modeling technique. Nevertheless the basic difference between stocks and flows is very important outside SD modeling, too. Let me give two examples taken from economic life:

Public debt vs. federal deficit:

In Austria, the public debt has become a serious problem over the last 30 years. Although one of the wealthiest nations in the world, this small nation with 8 million inhabitants accumulated a public debt of about 200 billion Euro up to the year 2000.[2] Nevertheless it is not the public debt, but just the federal deficit, i.e. the annual increase in public debt, which has been considered in the public political debate in Austria. Politicians claimed a decrease of federal deficit from say six billion Euro in the year 1996 to four billion Euro in the year 1997 as a "budgetary consolidation". In fact, nothing has been consolidated, the public debt increased in both years together by another ten billion Euro.

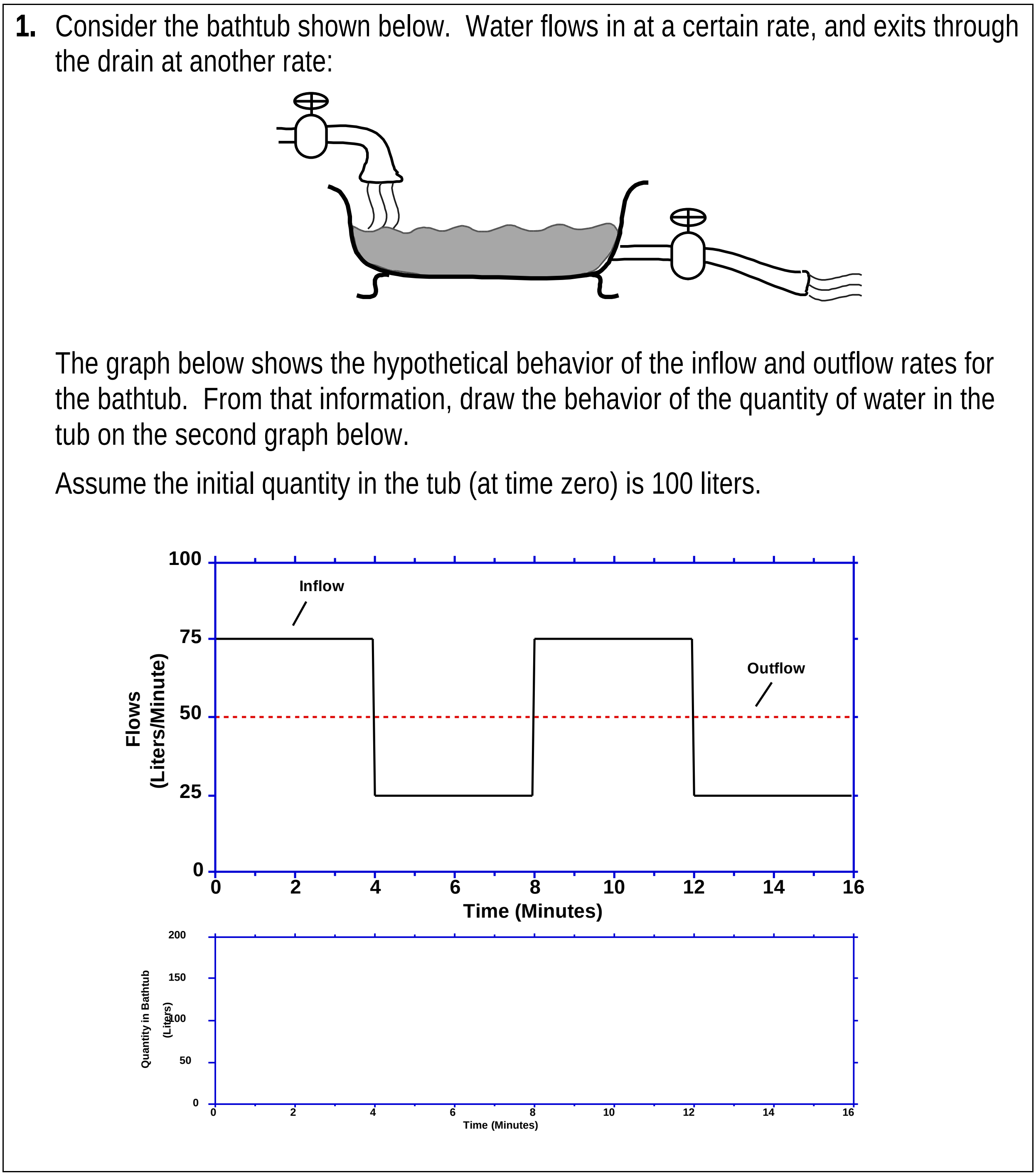
[1] Different aspects of systems thinking are discussed in some detail in Dörner (1996), Draper (1993), Forrester (1994), Frensch/Funke (1995), Gould (1993), Gomez/Probst (1997), Richardson (1991), Richmond (1991, 1993), Senge (1990) or Vester (1999). For a comparative overview concerning different uses of the term "systems thinking" see Ossimitz (2000, pp 9-62).

[2] This is about half the total public debt all 52 states of the whole African continent – which has a 100 times the population of Austria – had at the beginning of year 2000.

Balance sheet vs. profit-and-loss-statement:
I occasionally ask business administration students in my lectures about the difference between the figures in a balance sheet and the corresponding profit-loss-statement. Of course they know which figures are to be found in each of these sheets, but they find it very difficult to say *why* these data are organized in two different sheets and are not combined in a single sheet. Almost no one recognizes that the figures in a balance sheet refer to a point of time, e.g. Dec 31st 2000, whereas the figures in a profit-and-loss-statement refer to a time period, e.g. the whole year 2000. This means the balance sheet contains only stocks, the profit-and-loss-statement just flows.

2 The Sweeney/Sterman (2000) Study

My own study was motivated by John D. Sterman's plenary presentation of the Sweeney/Sterman (2000) paper at the 2000 SD Conference in Bergen, Norway. There John Sterman presented primarily the results of two "bathtub dynamics" tasks, which were presented with two different cover-stories (the literal filling of a bathtub, see Fig. 1, and a more abstract cash-flow.).



1. Consider the bathtub shown below. Water flows in at a certain rate, and exits through the drain at another rate:

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.

Fig. 1. Original "Bathtub-Task 1" from Sweeney/Sterman (2000)

Sterman stressed in his plenary presentation that the results of the (highly educated) MIT-students were rather poor. Sweeney and Sterman had identified seven criteria which were then used to evaluate each drawing. Table 1 gives the percentage of correct answers according to the different criteria.

		Swweney/Sterman N=95	Ossimitz N=154
31	When the inflow exceeds the outflow, the stock is rising	0,87	0,42
32	When the outflow exceeds the inflow, the stock is falling	0,86	0,43
33	The stock should not show any discontinuous jumps (it is piecewise continuous)	0,96	0,64
34	The peaks and troughs of the stock occur when the net flow crosses zero.	0,89	0,56
35	During each segment the net flow is constant so the stock must rise (fall) linearly.	0,84	0,38
36	The slope of the stock during each segment is +/- 25 units/time period.	0,73	0,26
37	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 units.	0,68	0,27
	Mean for all items	**0,83**	**0,42**

Table 1 Performance on the Bathtub Task 1.

In my own study there were 154 subjects, divided into three groups:

- 104 students at the beginning of their business administration studies at Klagenfurt University (Masters degree). Their age was mostly 19-21.
- 25 students (from different Viennese universities and various departments) just before playing the beer-game. Their age varied from 20 – 30.
- 25 students of environmental systems studies at the beginning of a systems modeling course at the University of Graz. Their age varied from 19 – 24.

About 60% of the 154 subjects of my study had just finished their secondary school education at age 18 or 19. The rest were students of higher semesters at university, but without a university degree.

In general, the results of the bathtub – tasks in my own study were considerably poorer than those of the MIT study. For a more detailed discussion see 4.3 and 4.4.

3 My own Study

The Sweeney/Sterman (2000) study did not disclose whether the observed problems of the subjects were due to deficits in stock-flow-thinking or due to deficits in the ability of reading and interpreting graphs. So I tried to create some additional tasks which should help to discern between these issues. My own study contained six tasks:

(1) Federal Deficit vs National Debt (FDND-Task)
(2) Arrivals and departures in the Alpenhotel (ADA-Task)
(3) Bathtub Task 1 (from Sweeney/Sterman 2000, see Fig. 1) (BT1-Task)
(4) Bathtub Task 2 (from Sweeney/Sterman 2000)[3] (BT2-Task)
(5) Filling of an Oiltank (FO-Task)
(6) Filling and emptying of a Bathtub (FEB-Task)

The *Federal Deficit vs. National Debt Task* (FDND) intended to test whether the subjects were aware of the difference between national debt (as a stock) and federal deficit (as a net flow). No graphs were involved in this task.

The *Arrivals and Departures in the Alpenhotel Task* (ADA – see Fig. 2) was designed to check the ability of the subjects to read and interpret a graph showing the number of guests arriving and departing on a daily basis. In this graph two flows were given, and qualitative answers concerning the stocks behind the flows had to be answered.

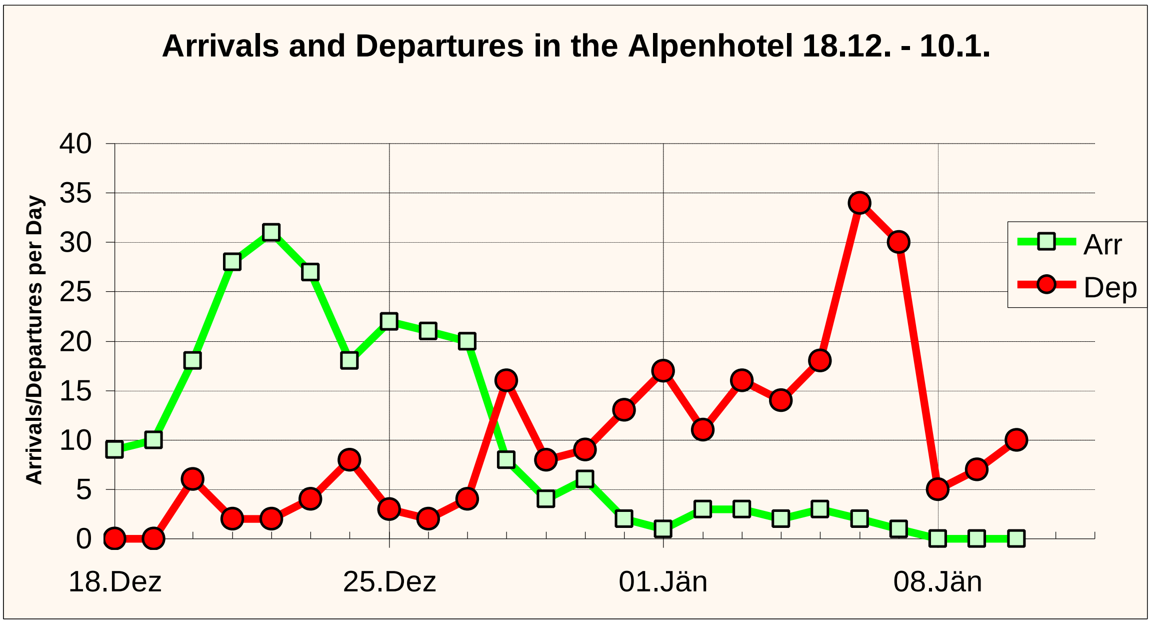


Fig. 2: Arrivals and Departures at the Alpenhotel (ADA)

[3] In this task the inflow followed a sawtooth wave, decreasing and increasing linearly between +100 and 0.

The bathtub tasks BT1 and BT2 should help to compare the achievements of the subjects of my own study with those of the Sweeney/Sterman study. The subjects were given exactly the same tasks as by Sweeney/Sterman (2000)[4].

The *Oiltank-Filling-Task* (OF – see Fig. 3) was designed to interpret a flow graph which is even simpler than the Bathtub task BT1. The inflow of oil starts at t=1 instantly with a rate of 200 Litres/Minute and stops at t=16. The questions should indicate the ability of the subjects to interpret the graph correctly.

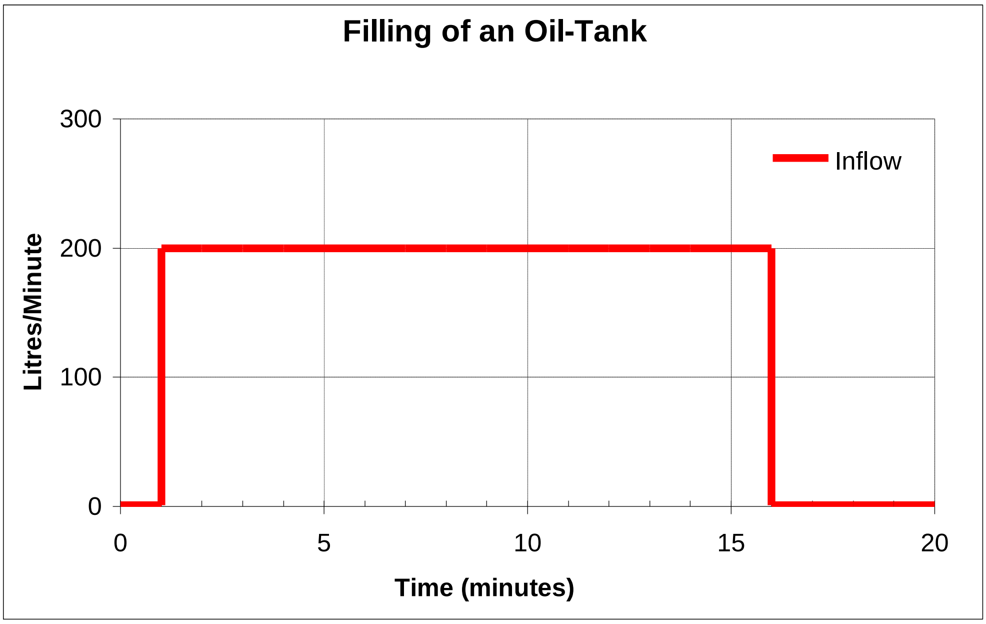


Fig. 3: Oiltank-Filling-Task (OF)

The *Filling and Emptying of a Bathtub Task* (FEB) was intended to check the ability to create a graph of the amount of water in a bathtub according to a story given in a reading text. This was designed to help to identify stock-flow-thinking capabilities without interference from the subjects' ability in interpreting graphs.

4 Basic Results

The study was undertaken with 154 students (62 female, 92 male). Most of the volunteers were born between 1979 and 1981 (=aged about 20 years at the time of the investigation in autumn 2000). Sixty of the subjects stated subjectively that they had excellent or good experience in reading and interpreting graphs,

[4] I translated the text into German; the graphics were simply copied from Sweeney/Sterman without alteration. In the text the terms "inflow" and "outflow" were translated as "Zufluss" and "Abfluss".

only eleven voulunteers thought that their ability to deal with graphs was poor or very poor. About half of the persons stated that their mathematics grade at the final "Matura"[5] exam was excellent or good, which is higher than the average of all mathematics grades at Matura exams in Austria.

The last preliminary question "Have you ever heard about stocks ("Bestandsmassen") and flows ("Bewegungsmassen") in statistics?" received the answer "no" from 83% of the test candidates.

4.1 The Federal Deficit vs National Debt – Task (FDND)

The aim of the FDND task was mainly to reveal how the subjects were able to discern between National Debt as a stock and Federal Budgetary Deficit as the corresponding net flow.

In a land called Fantasia the amount, by which the federal expenses exeed the federal income in one year is called the "Federal Budget Deficit". In 1998 the Federal Budget Deficit in Fantasia was 60 Billion Dollars, one year later it was 40 Billion Dollars. Please check which of the following statements are either right, wrong or not answerable! If you are not sure, please check don't know!

Fig. 4 Cover-Story for the FDND task

The following table shows the distribution of the answers to the individual questions[6].

		right	wrong	not ans-werable	don't know
10	In the year 1999 20 Billion Dollars of public debt have been paid back	28%	68% (m: **76%** / f: **56%**)		3%
11	The Minister of Finance could reduce the public debt from 1998 to1999 by a third.	53%	36% (**45**% / **24**%)	8%	3%
12	If the Minister of Finance in Fantasia is able to reduce the federal budget deficit to zero Dollars, (a balanced budget), then Fantasia does not have any more debt.	19%	62% (70% / 52%)	13%	6%
13	The public debt in Fantasia grew both in 1998 and in 1999.	44% (**60%** / **21%**)	29%	19%	6%
14	If the Minister of Finance in Fantasia is able to reduce the federal budget deficit to zero Dollars, (to budget balanced), then Fantasia has reached its highest public debt ever.	17% (**24**% / **6%**)	52%	21%	9%
15	A decrease in federal budget deficit implies automatically a decrease in the public debt.	44%	42% (**55**% / **23**%)	7%	6%

Table 2: Performance on the FDND task

The percentages in the second rows in the shaded cells give the percentage of correct answers among the male (left) and female (right) subjects. Any of the

[5] The Austrian "Matura" is equivalent to the German "Abitur" exam. It finishes secondary education after 12th or 13th class and qualifies for any university study in Austria. Its level is somewhere between the US high-school-degree and a first college degree.

[6] The boxes marked grey represent the correct answers. For item 10 both answers "wrong" and "not answerable" were considered correct.

indicated differences is significant on a 5% level[7], the bold printed values are significant on a 1% error level.

The average performance on each item was less than 45%! To make clear how disastrous this result is we should remind ourselves of the following: If I had instructed 70% of 154 complete ignoramuses to throw a coin to choose by chance between "right" and "wrong" and the other 30% to roll some dice to choose with equal chance one of the three answers "right", "wrong" or "not answerable", the expected average performance of this club of idiots flipping coins and rolling dice would also be 45%!

Out of the 45% performance the males consistently scored significantly higher than the females. We will see that this is true also for most of the other tasks. For me this finding is a big surprise. It definitely cannot be explained simply by saying that the female subjects are generally inferior to their male colleagues. The average math grades at the Matura of the female students was 1,9 (on a scale of 1(best) – 4(worst), whereas the men scored just 1,97 on average. Most of the students attended my own lessons after the test and also here I did not find that the females were inferior in their achievements compared to the males.

Although I have considerable experience in empirical studies in a broad variety of educational issues, this is actually the very first time that some significant difference in the performance between the sexes appeared.

4.2 Arrivals and departures in the Alpenhotel – task (ADA)

In the ADA-task the candidates were given an inflow and an outflow (arrivals and departures) of hotel guests as a graph (Fig. 2). The main goal of this task was to see whether the subjects can infer the stock of guests from these flows. The task was constructed in such a way that in the first period from Dec 18th until Dec 27th every day the number of guests arriving was higher than the number of guests departing. For the rest of the time the opposite was true. So from Dec 18th until 27th the number of guests was rising and from Dec 28th until Jan 10th the number of guests was decreasing. So the maximum number of guests was in the hotel on the night from Dec 27th to Dec 28th. In questions 21 and 22 this fact was addressed. The results were disastrous.

21) How can the graphic be used to discern (in an fast and elegant way, without any tedious calculations, but just by looking at the graph!) when the most guests were in the hotel? Explain how you might manage this.

[7] The hypothesis that both sexes have the same percentage of correct answers was tested using a Chi-Square independence test.

Only 16% of the subjects gave an answer which could be interpreted as something like "where the two lines intersect". 27% of the answers could be put in a group where the subject stated at least that the arrivals and departures had to be compared and the difference between these values is somehow relevant. 10% gave plainly the wrong answer "where the number of arrivals is maximum". Most of the verbal answers to this item were messy and hard to interpret.

22) On which day were the maximum number of guests in the Hotel ?

This item was correctly answered by just 24% (m: **34**%, f: **10**%) of the subjects[8]. 60% (m: **48**% f: **74**%) of the answers were Dec 22nd or one day before or after that day, which is the peak of the arrivals. Another 9% of the subjects took the peak of the departures (Jan 6th ±1 day) as their answer.

Item 23 (On which day was the maximum number of departures?) was correctly answered by virtually all subjects. Just 4% picked Jan 7th or 5th, while 3 out of 154 persons gave a completely wrong answer.

The comparision between the answers to items 22 and 23 allows the interpretation that subjects tend to interpret a graph as a graph of stocks, even if it is actually a graph of flows. To read the maximum from a curve in a graph seems to be a trivial task, if we ignore the minor problems of some subjects to read the horizontal time axis (which is a bit uncommon in this example) precisely. In item 22 a highly significant difference between the genders can again be observed.

[8] Any of the answers Dec 27th, Dec 28th or "the night from Dec 27th to 28th" was counted as correct.

4.3 Bathtub Task 1 (taken from Sweeney/Sterman 2000) – (BT1, see Fig. 1)

The bathtub tasks were taken with permission from the Sweeney/Sterman study. The main results in Table 2 show that the performance was considerably poorer than in Sweeney/Sterman (2000)[9].

		Swweney/ Sterman	Ossimitz
31	When the inflow exceeds the outflow, the stock is rising	0,87	0,42 (m: **0,53** / f:**0,24**)
32	When the outflow exceeds the inflow, the stock is falling	0,86	0,43 (m: **0,55** / f:**0,24**)
33	The stock should not show any discontinuous jumps (it is piecewise continouos)	0,96	0,64 (m: 0,72 / f:0,53)
34	The peaks and troughs of the stock occur when the net flow crosses zero.	0,89	0,56 (m: **0,66** / f:**0,40**)
35	During each segment the net flow is constant so the stock must rise (fall) linearly.	0,84	0,38 (m: **0,52** / f:**0,18**)
36	The slope of the stock during each segment is +/- 25 units/time period.	0,73	0,26 (m: **0,37** / f:**0,10**)
37	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 units.	0,68	0,27 (m: **0,37** / f:**0,11**)
	Mean for all items	**0,83**	**0,42** (m: **0,53** / f:**0,26**)

Table 3: Performance on the BT1

As in the other tasks, also in the BT task the responses of the female subjects are significantly weaker than the responses of the males: The average performance of the male subjects was 53%, about twice as high as that of the female subjects.

The criteria 31 to 37 are far from being independent of each other. On the contrary: the criteria 31 and 32 resp. 36 and 37 are so highly correlated with each other that each pair 31/32 and 36/37 can essentially be considered as a single variable. In both cases only a few outliers prevent a perfect correlation. Another very strong correlation is between the criteria 31/32 and 35: only 3 persons who miss criterion 31 succeed in criterion 35; just 8 of 95 persons who miss criterion 35 are successful in criterion 31.

For other pairs of criteria the correlation was just "one-sided": e.g. those who failed in criterion 31 trivially had no chance of succeeding in criterion 36. Out of the 64 who succeeded in 31, 40 were also successful concerning criterion 36. When comparing criteria 31 and 33 we have a somewhat different "one-sided" situation: of the 64 subjects who succeeded in criterion 31 only 2 failed in criterion 33. Very similar to the correlation between 31 and 33 is the correlation between 31 and 34.

What can be learned from the interrelations between the criteria? The most basic fact is that the ability to master criterion 31 *"when the net flow is positive, then the stock increases"* (and its counterpart 32) is crucial for the whole set of other criteria. Whoever fails on criterion 31, will most probably fail on all the others, too. It seems noteworthy to me that criterion 31/32 is not a quantitative, but

[9] The percentages would have been slightly better if the 16 of 154 subjects who have given no answer to the BT1 task at all had been discarded.

essentially a qualitative criterion: it argues just with the sign of the net flow and the resulting monotony of the stock. We can say that criterion 31/32 is the most fundamental and elementary aspect of any stock-flow relation. In other words, if this aspect is not understood properly, it is legitimate to say that the stock-flow-relationship is not understood at all. On the other hand, for those who do understand the difference between stocks and flows, criterion 31/32 is absolutely evident.

The main didactic conclusion that can be drawn from this analysis is very straightforward: teaching the fact that a positive net-flow induces an increase in the stock and vice versa might be the most fundamental lever for acquiring a practical knowledge of stock-flow relationships.

4.4 Bathtub Task 2 (from Sweeney/Sterman 2000)[10] – BT2

The BT2 task is structurally very similar to the BT1 task. The linear change of the flow induces a non-linear (actually parabolic) change of the stock. This makes this task considerably harder than BT1. This can be easily seen when the figures of table 4 are compared with those of table 3.

		Swweney/ Sterman N=79	Ossimitz N=154
41	When the inflow exceeds the outflow, the stock is rising	0,46	0,25 (m: **0,38** / f:**0,06**)
42	When the outflow exceeds the inflow, the stock is falling	0,41	0,23 (m: **0,36** / f:**0,03**)
43	The stock should not show any discontinuous jumps (it is piecewise continouos)	0,99	0,71 (m: 0,74 / f:0,66)
44	The peaks and troughs of the stock occur when the net flow crosses zero.	0,41	0,23 (m: **0,36** / f:**0,03**)
45	The slope of the stock at any time is the net rate.	0,25	0,10 (m: 0,13 / f:0,06)
46	The quantity added to (removed from) the stock during each segment of 2 time units is 50 units. The stock therefore peaks at 150 units and reaches a minimum of 50 units.	0,34	0,14 (m: **0,37** / f:**0,10**)
	Mean for all items	**0,48**	**0,28** (m: **0,39** / f:**0,16**)

Table 4: Performance on the BT2

Concerning the gender-specific differences, the gap between the performance of the male and femals subjects was even bigger for the BT2 task. Items 42 and 44 were extreme: almost all of the persons who succeded on these criteria were male!

As in BT1 task, the criteria of the BT2 task are highly correlated. The basic pattern is the same as in the BT1 task:

- Items 41 and 42 are in almost perfect correlation

[10] In this task the inflow followed a sawtooth wave, decreasing and increasing linearly between +100 and 0.

- A subject who fails on criterion 41/42 most probably fails all the other criteria, too.

4.5 *Filling of an Oiltank* (FO)

The FO task consisted of two parts. In the first part the discontinuities in the graph should be explained. Two of three persons confirmed that such a graph might be realistic, 25% of all subjects thought it might be not. Their arguments differed highly in content and quality, so that it was very hard to group them. About 9% of all persons made the point that the vertical discontinuity of the graph would mean that the valves of the filling device had to open instantaneously. 6% argued that such an instant change in flow is impossible. Another 5% of all test-candidates argued explicitly in a way which made it clear that they thought that the graph resembles the stock of oil in the tank.

The variety of answers to item 51 made one thing clear: there are numerous ways that the graph could be understood in a different way than was intended by me. One of the crucial aspects was the precise meaning of the word "filling". Some subjects inferred(rr-r?) from the facts that the graph was titled "Filling of an Oil-Tank" and that the time-axis was labeled from 0 – 20 minutes, that the filling lasted for 20 minutes, since the "whole graph" was titled "Filling of an Oil-Tank". Some of them argued that in the first and last four minutes the whole thing was just connected, which has to be counted technically as "filling" time, but no oil was flowing. Others argued that in the first and last four minutes the thick inflow-line is just a trifle above the horizontal zero line, so that a minimum of flow took place during these periods. I was baffled by how many different plausible interpretations could be given for the same graph! These ambiguities have to be taken into account when interpreting items 53, 55 and 56.
The following table gives the percentages of answers for all items[11].

Check the appropriate boxes!

		right	wrong	not answerable	don't know
52	The oiltank has been filled 200 cm high with oil.	8%	38%	47%	3%
53	The filling lasted 16 minutes.	21%	71%	4%	1%
54	Altogether 200 Liters have been filled in.	18%	76% (m: 83% f: 66%)	4%	1%
55	After 16 Minutes 200 litres have been let out of the tank.	14%	68%	13%	3%
56	The filling lasted 15 minutes.	68%	27%	2%	1%
57	The oiltank's capacity is 200 litres maximum.	4%	74%	16%	4%
58	After 16 minutes there were 3000 litres more in the tank than before.	64% (m: **74**% f: **50**%)	22%	3%	7%

Table 5: FO –Task (correct answers are shadowed)

[11] The balance to 100% did not respond.

The performance on the OF-task was clearly better than on any other task of my study. For any item the percentage of correct answers was about 64 – 85%.[12]

For items 53 and 56 I did not highlight a "correct" answer, because of the different possible meanings of the word "filling", which has crucial influence upon the question of which answer is correct. An analysis of the cross-tabulation of the answers to these two items revealed some interesting details. No less than 14 out of 154 persons checked both assertions 53 and 56 as "wrong". Those could be from the fraction who believes that the whole filling lasted for 20 minutes. But even more astonishingly 9 persons actually checked both assertions as "true". Maybe they interpreted the statement "lasted 15 minutes" as "lasted at least 15 minutes" and not as "lasted exactly 15 minutes".

The results of the FO – Task were less gender-specific than for the other tasks. A highly significant difference[13] could be observed just for the percentage of correct scores on item 58 (m: 74%, f: 50%); another significant difference[14] could be found for the distribution of correct answers for item 54.

I also tried to find some correlations between the performance of the FO – Task and the ADA – Task. Both tasks are based on the correct interpretation of a flow graph. At the present state of my investigations no positive correlations can be reported. The only anomalies I found were rather strange. When comparing the scores of items 22 ("On which day are most guests in the Alpenhotel?") and 64 it turned out that 36 of the 38 persons who gave the definitively wrong answer Jan 6th (or Jan 5th or 7th) to item 22 answered item 64 correctly as "wrong"; the other two gave the answer "not answerable" to item 64.

In short, I can summarize that the FO task results seem to be somewhat independent of the results of the other tasks.

4.6 Filling and Emptying of a Bathtub (FEB)

The FEB task should test the ability of the subjects to transpose a story about the filling and emptying of a bathtub into a graph of the amount of water over time. The following table gives an overview of the results.

[12] For Items 52 and 55 both answers "wrong" and "not answerable" were generously counted as correct.
[13] Error-Level for Chi-Square: better than 1%.
[14] Error-Level for Chi-Square: between 1 and 5 Percent.

FEB-Task scores	abs	all	male	female
Correct solution	44	29%	**38,0%**	**14,5%**
Almost correct, minor inaccuracies	39	25%	27,2%	22,6%
In the first 4 minutes the amount of water was considered to be constant, otherwise correct	10	6%	6,5%	6,5%
Partial solution or serious errors, but somehow plausible in the overall picture	19	12%	9,8%	16,1%
Very severe errors, very poor solution or just a minor part of the graph sketched	36	23%	**15,2%**	**35,5%**
No solution	6	4%	3,3%	4,8%
N=	**154**	100%	100,0%	100,0%

Table 5: FEB-Task

In the FEB task slightly more than half of all the answers were correct or at least almost correct. We again see a significant gender trend: among the correct solutions male subjects are overrepresented, while in the category of the poorest solutions the women are overrepresented.

An interesting question was whether the performance on the FEB task is somehow correlated with some other score. First of all I could find a positive correlation between the self-estimation of the subjects about their ability to deal with graphs and with the performance on the FEB – Task. Of the nine persons who estimated themselves as very good in dealing with graphs, four did indeed provide a perfect solution to the FEB-Task. Similarly four of the six subjects who estimated their knowledge concerning graphs as very poor were scored in the poorest category on the FEB task.

A comparision between the FEB-Task and the BDND and ADA – Tasks revealed some partial trends, but no result with statistical significance. A curious detail is that almost all subjects who chose the wrong answer Jan 6th on item 22 performed well or very well on the FEB Task.

The cross-tabulation of the FEB-Scores with the "basic" criterions 31 and 41 *"when the inflow exceeds the outflow, the stock is rising"* of the BT1 and BT2 tasks brought significant results. The subjects who fulfilled criterion 31 were significantly better[15] in their performance of the FEB-Task. The correlation between the criterion 41 and the FEB-Scores was highly significant on the 1%-level. This gives some indication about a certain relationship between the BT and BEF tasks.

5 Further investigations

At the present stage of the evaluation of the results only a few of the many possible correlations between the test items have been checked for significance. One of the very clear findings of this survey is the massive gender-bias of the

[15] The error-level is only slightly below 5%.

results. I sincerely must admit that I have no explanation and not even a serious hypothesis about the reasons for this finding.

Another interesting result emerges from the high correlation between several criteria of the BT1 and BT2 tasks. These correlations indicate that there might be a fundamental aspect of stock-flow-thinking, which is a prerequisite for more elaborate stock-flow-thinking capabilities. This fundamental aspect is the ability to grasp that in a stock-flow-context the stock with one inflow and one outflow is increasing when the inflow is bigger than the outflow (or the net flow is positive, to put it in another way). Some findings indicate that this might be a key criterion for discriminating between a stock-flow-thinker and a non-stock-flow-thinker.

The present study raises a number of puzzling questions and indicates the need for further investigations. The most serious issue is the fact that the practical abilities in stock-flow-thinking and in interpreting graphs among students at different Austrian Universities are shockingly poor. The situation among the Austrian students is much severer than that of the highly educated MIT-students, which Sweeney and Sterman described as being "on a poor level" (Sweeney/Sterman 2000).

I hope that a closer investigation of the material of this study can reveal some additional clues about how to deal with this problem in education. It is my intention to develop within the next months a "Stock-Flow-Thinking Crash Course", which is designed to overcome within a few hours of instruction the deficits that have been shown in the Sweeney/Sterman and in my own study. The crash-course should be applicable for a very broad audience; without many technical requirements in mathematics, computer science or SD modelling.

I plan to evaluate the efficiency of this crash course in a pre-test – post-test design. The pre-test should take place one or two months before the course; the post-test at least two months after the course. In this way I hope to gain some insights into the long-term efficiency of the course.

References

Dörner, Dietrich (1996): The Logic of Failure. (dt. Die Logik des Misslingens.) New York: Metropolitan Books/Henry Holt

Draper, Frank (1993): Aproposed sequence for developing systems thinking skills A proposed sequence for developing systems thinking in a grades 4 – 12 curriculum. System Dynamics Review 9(2), 207-214.
URL: ftp://www.clexchange.org/documents/system-ed/SE1993-01AProposedSequence.pdf

Forrester, Jay W. (1961): Industrial Dynamics. Cambridge, MIT Press

Forrester, Jay W. (1994): System Dynamics, Systems thinking and Soft OR. System Dynamics Review 10(2)

Frensch, Peter A./ Joachim Funke (eds) (1995): Complex Problem Solving – The European Perspective. Mahwah: Lawrence Erlbaum Associates, Inc.

Gould, J. (ed) (1993): Systems Thinking in Education. Systems Thinking Review (special issue)

Gomez, Peter / Gilbert J. Probst (1997): Die Praxis des ganzheitlichen Problemlösens: vernetzt denken, unternehmerisch handeln, persönlich überzeugen. Bern: Paul Haupt

Ossimitz, Günther (2000): Entwicklung systemischen Denkens. München: Profil Verlag

Richardson, George (1991): Feedback Thought in Social Science and Systems Theory. Philadelphia: University of Pennsylvania Press

Richmond, Barry (1991): Systems Thinking: Four Key Questions. Lyme: HPS Inc.

Richmond, Barry (1993): Systems Thinking: Critical Thinking Skills for the 1990ies and beyond. System Dynamics Review 9(2), 113 – 133.

Senge, Peter (1990): The Fifth Discipline: The Art and Practice of the Learning Organization. New York: Doubleday.

Sweeney, Linda Booth / John D. Sterman (2000): Bathtub Dynamics: Preliminary Results of a Systems Thinking Inventory

Sterman, John D. (2000): Business Dynamics. Systems Thinking and Modeling for a Complex World. New York: Irwin/McGraw-Hill

Vester, Frederic (1999): Die Kunst, vernetzt zu denken. Stuttgart: DVA

Appendix A: Original testing form (translated)

4 min

1 Code Name:____________________

2 Year of Birth:_________ 3 Sex: male O female O

4 How well do you think you are able to read and understand graphs?

very well average very poorly
O O O O O

5 Mathematics grade at "Matura"-exam 1 – 2 – 3 – 4 – no grade

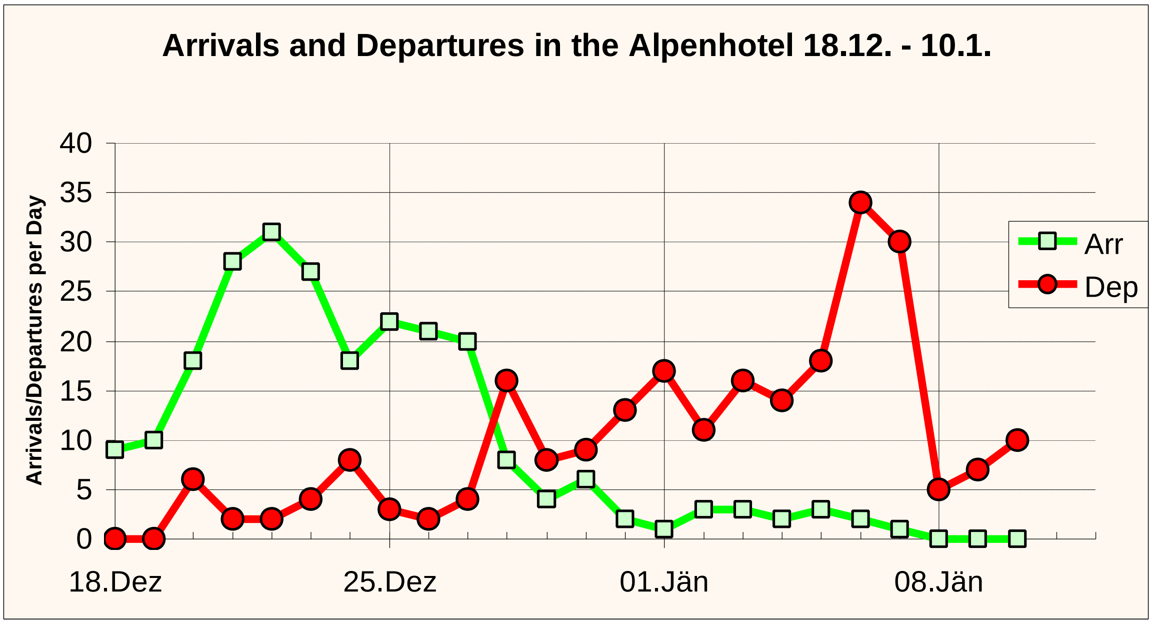
6 In Statistics Classes, have you ever heard anything about stocks (Bestandsmassen) versus flows (Bewegungsmassen)? Yes O No O

If yes, please explain the difference between stocks and flows.

1)
In a land called Fantasia the amount, by which the federal expenses exeed the federal income in one year is called the "Federal Budget Deficit". In 1998 the Federal Budget Deficit in Fantasia was 60 Billion Dollars, one year later it was 40 Billion Dollars. Please check which of the following statements are either right, wrong or not answerable! If you are not sure, please check don't know!

		right	wrong	not answerable	don't know
10	In the year 1999 20 Billion Dollars of public debt were paid back				
11	The Minister of Finance could reduce the public debt from 1998 to1999 by a third.				
12	If the Minister of Finance in Fantasia is able to reduce the federal budget deficit to zero Dollars, (a balanced budget), then Fantasia does not have any more debt.				
13	The public debt in Fantasia grew both in 1998 and in 1999.				
14	If the Minister of Finance in Fantasia is able to reduce the federal budget deficit to zero Dollars, (to budget balanced), then Fantasia has reached its highest public debt ever.				
15	A decrease in federal budget deficit implies automatically a decrease in the public debt.				

2) 2 min



The Alpenhotel opens for its Christmas season on Dec. 18th and closes on January the 10th. The graphic above shows for each day the number of arriving (light squares) and departing (dark circles) guests. Please answer the following questions:

21) How can the graphic be used to discern (in an fast and elegant way, without any tedious calculations, but just by looking at the graph!) when the most guests were in the hotel? Explain how you might manage this.

22) On which day were the maximum number of guests in the Hotel?

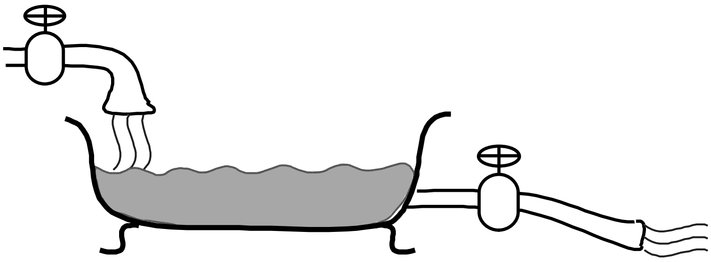
____

23) On which day there were the most departures? ___________

24) How many guests were in the Hotel after Jan 10th? __________

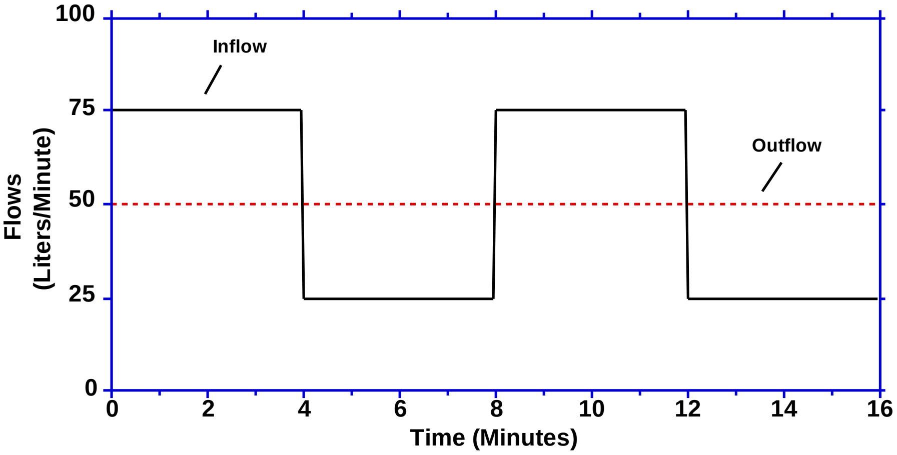
25) According to which criterion might you discern whether on a certain day the number of guests is increasing or decreasing?

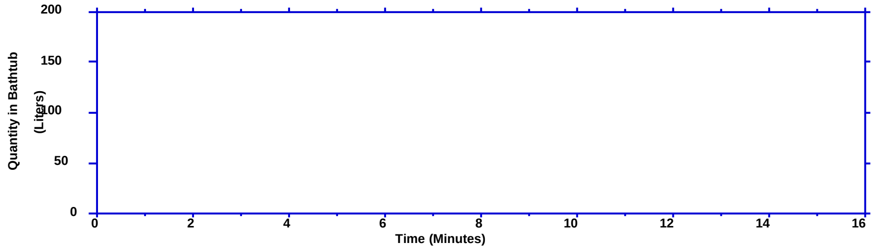
4)[16] **(4min)** Consider the bathtub shown below. Water flows in at a certain rate, and exits through the drain at another rate:



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.

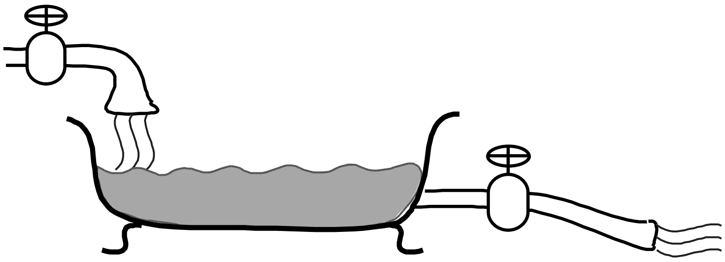
Assume the initial quantity in the tub (at time zero) is 100 liters.





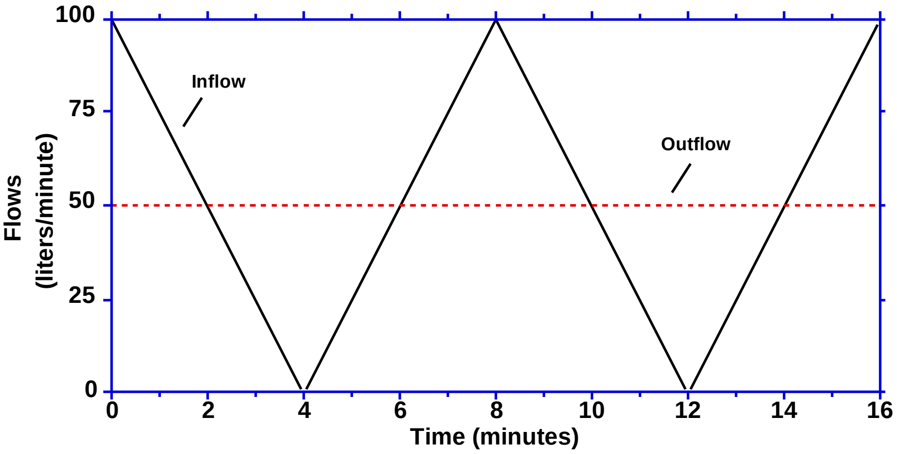
[16] This is exactly the same as bathtub task 1 of the Sweeney/Sterman.(2000) study.

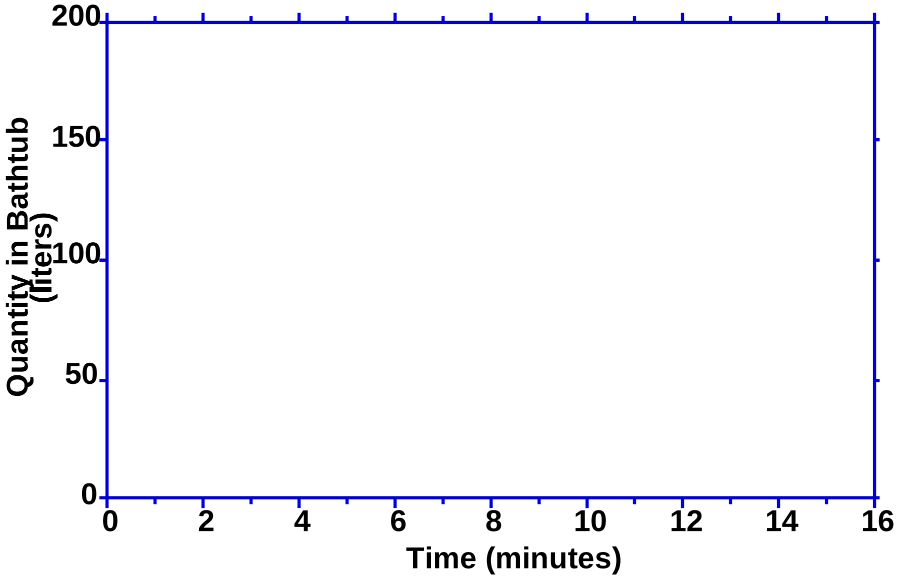
5)[17] **(4min)** Consider the bathtub shown below. Water flows in at a certain rate, and exits through the drain at another rate:



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.

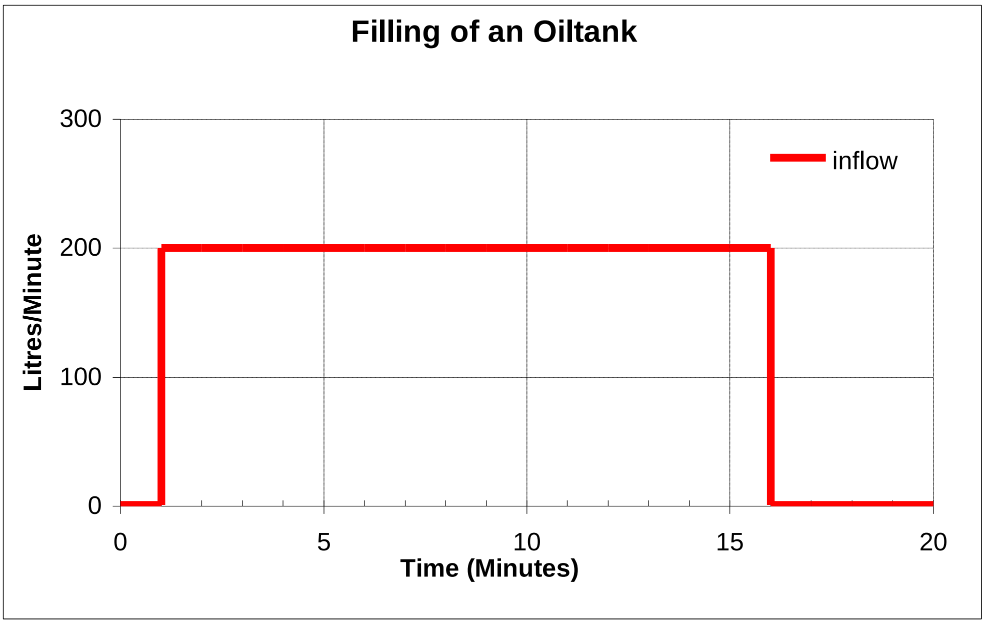
Assume the initial quantity in the tub (at time zero) is 100 liters.





[17] This is exactly the same as bathtub task 2 of the Sweeney/Sterman (2000) study.

5) (4min)
The following graph describes the filling of an oil-tank:



Please answer the following questions:

51) Are the discontinuities in the graph of the function at time 1 resp. 16 reasonably explainable? Yes O No O
If yes, how?____

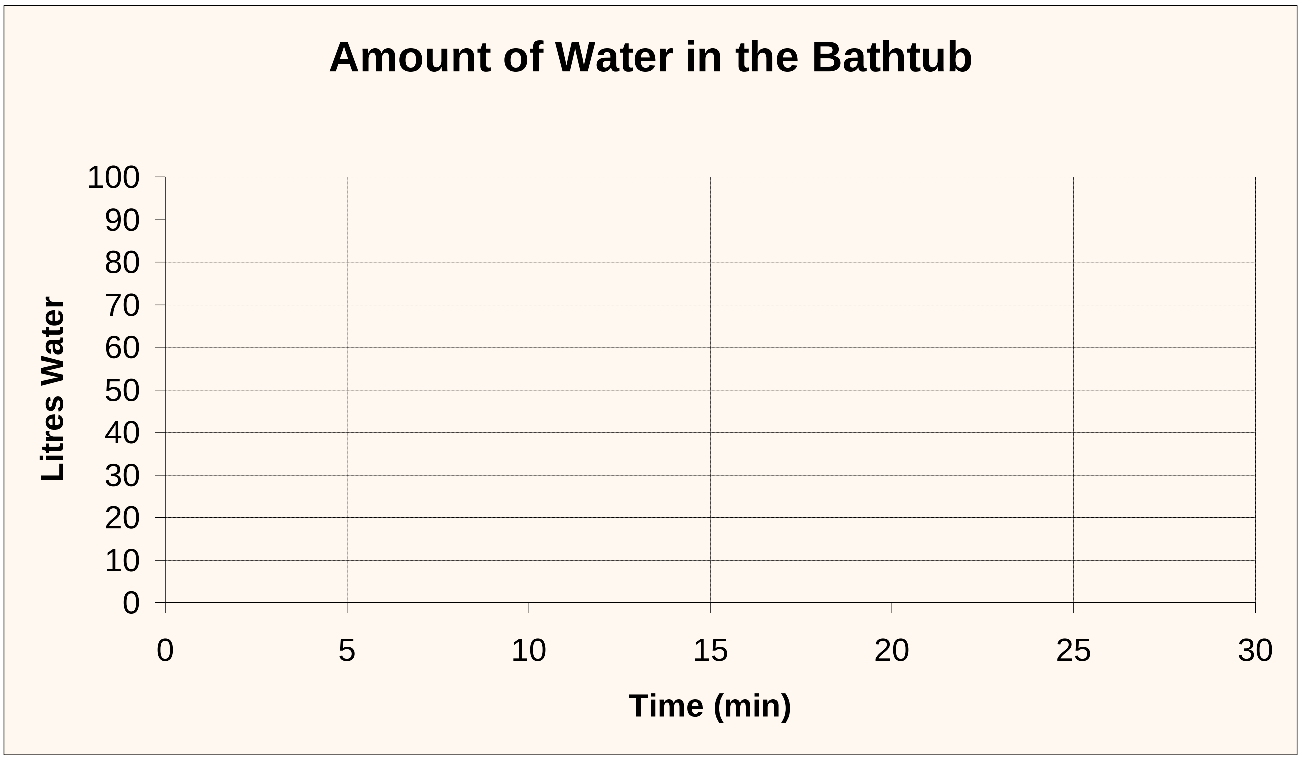
If no, why not?____

Which of the following statements are true, false or not answerable?

Check the appropriate boxes!

		right	wrong	not answerable	don't know
52	The oiltank has been filled 200 cm high with oil.				
53	The filling lasted 16 minutes.				
54	Altogether 200 litres have been filled in.				
55	After 16 minutes 200 litres have been let out of the tank.				
56	The filling lasted 15 minutes.				
57	The oiltank's capacity is 200 Liters maximum.				
58	After 16 minutes there were 3000 liters more in the tank than before.				

6) (3 min)
At 7:00pm exactly Mr. Maier starts to fill his (empty) bathtub. The inflow is constant at 14 liters/min. At exactly 7:04pm Mr. Maier notices that the outflow of his bathtub has been left open and he closes it. Through the open outflow exactly 9 Litres/min leave the bathtub. At excatly 7:09 Mr. Maier closes the inflow and enjoys his bath until 7:15pm. At exactly 7:15 he opens the outflow and lets all the water flow out.



61) Give in the above graphic an approximate sketch how the amount of water in the Bathtub will develop over the time!