Stock-flow failure can be explained by the task format

Helen Fischer & Christina Degen
University of Heidelberg

Abstract

Stock and flow (SF) problems are ubiquitous in nature, ranging from filling water into a tub to the accumulation of atmospheric CO_2. Research on the "SF failure" suggested, however, that people have severe difficulties understanding basic SF problems. We present the results of an experiment (N = 277) were participants solved a range of SF problems with varying task formats and semantic embeddings. Results indicate that (a) SF failure can at least partially be attributed to specifics of the task format used previously; (b) significant reductions in error rates can be achieved by only slight changes in the task format; and (c) a fundamental misunderstanding in the construction of graphs can explain a typical mistake in these tasks. The majority of participants arrived at the correct solution when SF problems were presented verbally. Implications for risk communication are discussed.

Keywords: Stock and flow problems; stock and flow failure; task format; climate change

Stock-flow failure can be explained by the task format

It is a well-established finding that humans have severe difficulties understanding stock-flow (SF) dynamics (see Sterman, 2011 for a recent review). Such dynamics typically comprise a stock, which accumulates over time and is dependent on a given in- and outflow progression. SF dynamics are pervasive in many areas of life, ranging from everyday phenomena such as the accumulation of money on a bank account or the regulation of body weight, to more abstract scenarios such as the supply line of a factory or the accumulation of CO_2 in the atmosphere. Even though most of these problems can contain multiple in- and outflows, the underlying principle is always simple and is often explained with the bathtub analogy, according to which the water level (stock) in a bathtub increases if the inflow of water through the faucet exceeds the outflow through the drain, and contrariwise drops if the outflow exceeds the inflow. Given the simplicity and ubiquity of SF dynamics, previous findings showing that graduate students at MIT had severe problems to solve even simple SF tasks, seem perplexing (Booth Sweeney & Sterman, 2000; Cronin & Gonzalez, 2007; Sterman & Booth Sweeney, 2002, 2007). It was concluded that humans lack understanding of SF dynamics, a phenomenon termed *SF failure* (Booth Sweeney & Sterman, 2000).

In this paper, we argue, however, that SF failure can at least partly be attributed to specifics of the task formats used in previous research. The experiment depicted here aimed at delineating difficulties caused by the task format (method of information display and required answer format) from the presumably inherent difficulties people have with understanding SF dynamics.

Research on SF problems

In the original paradigm investigating participants' understanding of SF dynamics (Sterman & Booth Sweeney, 2002, 2007), participants were typically presented with an introduction on the relationship between CO_2 emissions, absorptions (CO_2 taken up by biomass and oceans), and atmospheric CO_2 concentration. They were then presented with a graph depicting atmospheric CO_2 concentration (stock) stabilizing from the year 2100 onwards and with a graph depicting previous CO_2 emissions and absorptions. Participants were asked to sketch emission and absorption trajectories in

such a way that a stabilizing CO_2 concentration could be achieved. A repeated finding was that participants made use of a pattern matching heuristic, sketching in- and outflows that followed the trajectory of the stock, i.e., a continuous increase followed by stabilization. That way, drawn emissions typically exceeded absorptions leading to an actual increase of atmospheric CO_2 (Sterman & Booth Sweeney, 2002, 2007). In this original paradigm, SF failure was also demonstrated for multiple choice answer formats (e.g., CO_2 emissions resulting from human activity would have to: Gradually rise about 8% and then stabilize by the year 2100), different outcome scenarios (atmospheric CO_2 concentration stabilizing, CO_2 emissions dropping to zero, CO_2 emissions stabilizing), and different semantic embeddings including more familiar contexts than atmospheric CO_2 concentration (Booth Sweeney & Sterman, 2000; Cronin & Gonzalez, 2007; Sterman & Booth Sweeney, 2002, 2007).

Comprehension of Task Formats vs. Comprehension of SF Dynamics

In order to convincingly establish the validity of the SF failure, SF tasks need to assess construct-relevant misunderstanding of SF problems, rather than construct-irrelevant problems with the specific task format. Previous studies demonstrated the importance of this dissociation by showing that displaying isomorphic tasks in different formats can have a dramatic impact on problem-solving performance, such as on the Wason selection task (Cheng & Holyoak, 1985), the Tower of Hanoi (Kotovsky, Hayes & Simon, 1985), deductive reasoning (O'Brien et al., 1990), distributed cognitive tasks (Zhang & Norman, 1994), mathematical problems (Bassok, 2001; Landy & Goldstone, 2007), line graph description (Xi, 2010), and inference from complex graphical displays (Hegarty, Canhan & Fabrikant, 2010; Novick & Catley, 2007).

However, task formats in the original paradigm, in spite of variation, had one thing in common: a rather scientific notation. This notation included coordinate systems and graphs in both the information display and answer format and percentage values in the multiple choice answers. It was shown that comprehension of coordinate systems and graphs can be error-prone (Carpenter & Shah, 1998; Gattis & Holyoak, 1995; Shah & Carpenter, 1995) and that participants have difficulties dealing with percentage values (Gigerenzer & Hoffrage, 1995; Hoffrage & Gigerenzer 1998; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000). Consequently, we argue that the original paradigm might

have concealed participants' true understanding of SF dynamics, thus potentially causing construct-irrelevant variance.

Cronin and Gonzalez (2007) investigated the generalizability of SF failure in a series of experiments, varying the task context or the way in- and outflows were presented. However, in all tasks coordinate systems and graphs were used. Thus, potential problems with this aspect of the task format could not be eliminated. In a later experiment, Cronin, Gonzales and Sterman (2009) specifically investigated whether SF failure is a mere artifact of using coordinate systems by presenting participants with alternative display formats (line graphs, bar charts, texts, and tables; see Figure 1 for the textual display). The SF problem used, the so-called department store task, describes the number of people entering and leaving a department store over a period of time. Participants needed to determine at what time the most or fewest people were inside the store.

In a department store, people enter and leave over a 30-miunte period. In the first minute, 9 people enter and 8 leave. In the second minute, 10 people enter and 5 leave. In the third minute, 9 people enter and 8 leave. In the fourth minute, 14 people enter and 12 leave. In the fifth minute, 9 people enter and 8 leave. In the sixth minute, 9 people enter and 8 leave. In the seventh minute, 8 people enter and 8 leave. In the eighth minute, 7 people enter and 9 leave. In the ninth minute, 4 people enter and 13 leave. In the tenth minute, 7 people enter and 11 leave. In the eleventh minute, 10 people enter and 15 leave. In the twelfth minute, 8 people enter and 12 leave.

1. During which minute did most people enter the store?
2. During which minute did most people leave the store?
3. During which minute were the most people in the store?
4. During which minute were the fewest people in the store?

Figure 1. Textual display of the original paradigm (adopted from Cronin, Gonzalez & Sterman, 2009, p. 118).

To control for participants' comprehension of the information display, participants were asked at what time most people entered or left the department store.

Across all display formats, the majority of participants was able to answer these control questions correctly, but only the minority of the sample was able to draw correct inferences about the corresponding stock. The authors concluded that SF failure does not result from participants' inability to understand graphs, but rather from a fundamental error in human reasoning, i.e., SF failure.

However, this conclusion might be premature, for three reasons. First, difficulties with coordinate systems were not convincingly ruled out because for all display types, the control questions could be answered correctly by using simple salience heuristics picking the highest or lowest number displayed. Thus, no deep understanding of coordinate systems was necessary to answer the control question, but a deeper understanding of coordinate systems is arguably necessary to answer the SF tasks in the original paradigm. Second, the control questions only tested interpretation of graphs and not construction thereof, which was a prerequisite for solving tasks correctly in the original paradigm. Construction of graphs has been subject to little cognitive research, so that participants' ability to submit their answers by constructing graphs cannot necessarily be assumed. And third, in all data displays—even the textual—numerical information was salient (numbers of people entering and leaving). We argue that this salience of numerical information encourages participants to focus on and work with the given numbers, for example by performing simple calculations, rather than making an effort to detect the underlying SF structure. That is, the salience of quantitative information may encourage a focus on this specific information (local search) instead of an understanding of the deep structure (global search) (Guthrie, Weber & Kimmerly, 1993; Wainer, 1992). Of course, all SF tasks used previously did not have to be solved via calculating. However, due to the salience of numerical information, it may not have been clear to the participant that SF tasks can be solved with help of the problem structure beneath the numerical information. Therefore, we argue that in order to establish whether participants truly lack understanding of the SF structure, a more qualitative task format is needed.

Furthermore, the use of a pattern matching heuristic might have been encouraged by the original task format where the to-be-completed emissions followed the same trajectory as the resulting stock, i.e., a constant increase. In an experiment testing the generalizability of the pattern matching heuristic (Cronin, Gonzalez &

Sterman, 2009), participants were presented with graphs depicting various in- and outflow trajectories and needed to sketch the resulting stock. Even though in this task design no pattern was implied between inflows and the stock, participants adopted a pattern matching heuristic. We therefore want to find out whether the use of the pattern matching heuristic can be reduced when participants are presented with a strong hint against matching patterns: a mismatching pattern between inflow and stock. Presenting participants with the possibility that the stock increases even though the inflow decreases, for example, might enable participants to overcome the otherwise strong pattern matching heuristic and thus achieve higher solution rates.

To sum up, the present experiment investigated whether SF failure is a fundamental error in human reasoning and whether the pattern matching heuristic truly is a generic strategy. We argue that participants are capable of understanding SF structures and that SF failure can be attributed to the task format used previously, specifically to the use of coordinate systems and the salience of specific numerical information. The experiment thus tests whether participants answer SF tasks correctly when a purely verbal task format is used that encourages participants not to focus on and work with specific quantitative data, but to "get the qualitative gist" of the structural relationships depicted in the data (Shah & Hoeffner, 2002, p. 53).

Overview of the experiment

To test the validity of the SF failure, the study investigated whether different SF tasks measure construct-relevant aspects of the task (understanding of SF structure) vs. construct-irrelevant aspects of the task (understanding of the task format). To do so, three different types of SF tasks were constructed.

(1) Standard Tasks. Graphical and textual task formats used previously (Booth Sweeney & Sterman, 2000; Cronin & Gonzalez, 2007; Cronin, Gonzalez & Sterman, 2009; Sterman & Booth Sweeney, 2007) were administered with and without slight modifications. Modifications concerned whether additional numerical data was provided (initial stock or the exact outflow trajectory) and whether pattern matching was suggested in the task display (matching patterns of displayed inflow and stock trajectories). These modifications left the problem structure unaffected allowing us to test (a) whether the original paradigm encourages participants to focus on quantitative

surface features of the problem instead of trying to detect the problem structure. If this is the case, solution rates should vary as a function of the numerical information provided even though the problem structure remains constant. Furthermore, these modifications allowed us to test (b), whether participants use the pattern matching heuristic even when it is not suggested by the task format.

(2) Interpretation and Production Tasks (I/P tasks). I/P tasks examined the distinction between being able to interpret graphs versus being able to actively produce graphs and between submitting answers verbally versus graphically. These distinctions were introduced to shed some light on the little studied construction of graphs and potentially typical errors, and to investigate whether participants' potential understanding of SF dynamics is concealed in the original paradigm. If participants are able to answer SF questions correctly when submitting their answers verbally, but then make errors constructing the corresponding line graph, the original paradigm could not be seen as a fair test of participants' ability to understand SF problems.

(3) Verbal Tasks. Verbal tasks did not include coordinate systems or graphs, and little or no numerical information. Hence, verbal tasks tested whether SF failure could be reduced or even eliminated when an understanding of coordinate systems is not required, and when participants are encouraged to detect the qualitative gist of the problem structure.

We hypothesize the following:

(H1) Original task formats encourage participants to focus on numerical surface information instead of the underlying problem structure. Consequently, solution rates will vary as a function of the provision of additional numerical information. This applies even though the numerical information does not affect the problem structure.

(H2) The pattern matching heuristic is used more frequently in scenarios where pattern matching is suggested compared to scenarios where pattern matching is not suggested.

(H3) Even participants who correctly solve a given SF problem verbally may not be able to construct the according line graph into a coordinate system.

(H4) SF failure will be significantly reduced in a verbal task format without coordinate systems and with little or no quantitative information.

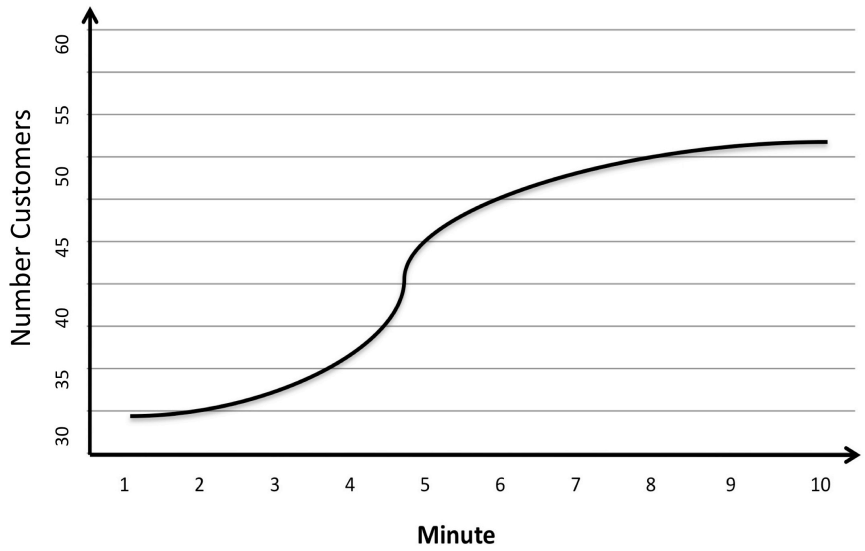
Method

Participants. A total of N = 277 participants (73% females) between 18 and 75 years of age took part in the experiment. Mean age was 34.33 (*SD* = 17.61). All participants gave written informed consent. The sample consisted of students from the University of Heidelberg and people from the general population recruited for a larger project. Participants received course credit or financial rewards (5 – 10 €, depending on decisions made in other parts of the study) for participation.

Materials. (1) Standard tasks: Standard tasks were adopted from previous research and were administered both as a graphical display (e.g., Sterman & Booth Sweeney, 2007) and as a textual display (Cronin, Gonzales & Sterman, 2009).

In the graphical display, three variations were developed that we will illustrate with the example of one scenario, the department store task (Fig.2):

The following Figure depicts how many customers are inside a department store over the course of 10 minutes:



The following Figure depicts the amount of people entering and leaving the above department store. Please draw both how many people must enter and how many

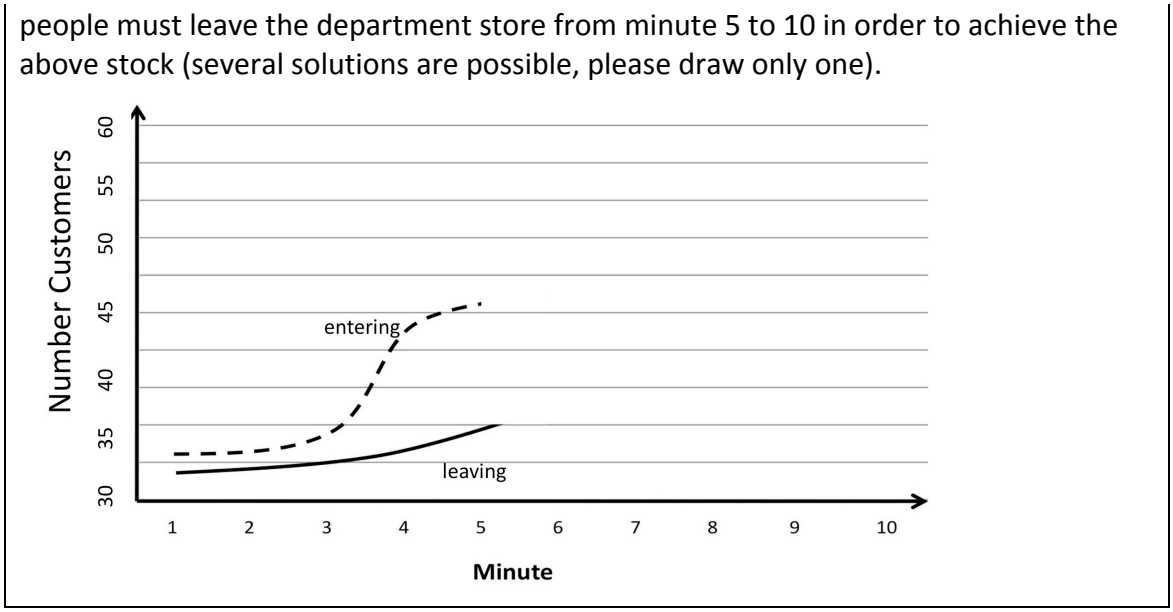


Figure 2. Example of the standard task (translated): department store scenario (translated). First, participants are presented with the development of a stock, then, participants are presented with the to-be completed in- and outflow.

The first variation refers to whether the initial stock (IS) is given with the problem display, or not (IS vs. ~IS condition). Figure 3 depicts the ~IS condition. In the IS condition, participants received the additional information (see italics): "The following figure depicts how many customers are inside a department store over the course of 10 minutes. *At the beginning, 32 customers are inside the store*". The second variation concerns the presentation of in- and outflow. While in the original paradigm initial inflow was depicted using a graph (line) and outflow was depicted using a single dot, we created an additional condition in which both the in– and outflow were presented as lines (2L vs. ~2L condition) delivering more numeric information about the exact outflow trajectory. Figure 3 depicts the 2L condition. In the ~2L condition, outflow was depicted using a single dot. IS and 2L conditions were fully crossed, resulting in four different tasks: ~IS, ~2L; ~IS, 2L; IS, ~2L; IS, 2L. Note that the ~IS, ~2L task is equivalent to the original task used in previous research and served as the baseline condition (Sterman & Booth Sweeney, 2002, 2007). These tasks comprised different scenarios: members of a club (~IS, ~2L), customers in a department store (IS, ~2L),

guests on a party (~IS, 2L), and nuts collected by a squirrel (IS, 2L), that varied not only in terms of the semantic embedding, but also in terms of the numbers used on the *x*- and *y*-axis in order to minimize carry-over effects (see Appendix I for a all scenarios used in the standard tasks).

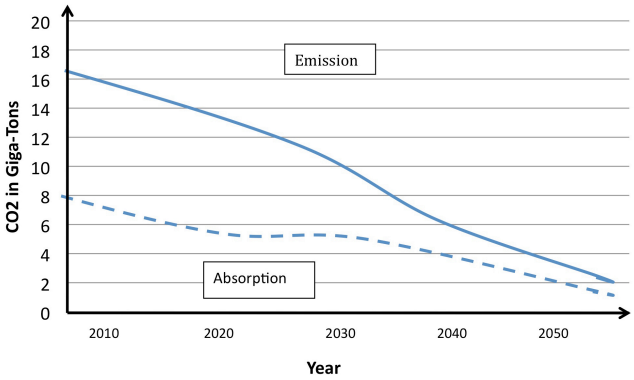
The third variation concerns the suggestion of pattern matching in the information display (PM vs. ~PM condition). In the PM condition, the inflow followed the same trajectory as the to-be completed stock, in the ~PM condition, the inflow followed a different trajectory as the to-be completed stock. Specifically, in the ~PM condition, the inflow followed a decreasing trajectory, while the stock followed an increasing trajectory; in the PM condition, both inflow and stock followed an increasing trajectory. (Note that neither 2L nor IS was varied over the two PM conditions, i.e., PM tasks can be understood as PM,~IS, ~2L and ~PM,~IS, ~2L.)

The textual display of the standard task was again adopted from previous work (department store task: Cronin, Gonzales & Sterman, 2009). Two variations of the original paradigm were realized: IS and ~IS. The ~IS condition was exactly the same as in the original task (Figure 2). In the IS condition, participants received the additional information (see italics): "The following text describes the amount of people entering and leaving a department store. *At the beginning, 32 people are inside the department store*."

(2) I/P Tasks: I/P tasks were again adopted from the original paradigm (e.g., Sterman & Booth Sweeney, 2007) and were administered in two scenarios (atmospheric CO_2 concentration, number of children on a playground). We will illustrate the tasks using the CO_2 scenario (Fig. 3). As in the original paradigm, participants first received a short introduction to the problem describing the relationship between CO_2 emissions, absorptions and atmospheric CO_2 concentration. Participants were then presented with a coordinate system depicting in- and outflows and four sub-tasks exploring fundamental understanding of the graphs (question 1), an estimate of the resulting stock (question 2), verbal production of necessary inflows and outflows given a decreasing stock (question 3), and the graphical production of the answer to question 3 into a coordinate system (question 4). Note that answers to question 3 (verbal production of in- and outflows) were deliberately simple in order to test whether participants necessarily produce a correct graphical answer given they are able to produce a correct verbal answer. The

structure of the playground scenario was the same, except for one difference: In question 3, participants were asked to name the necessary inflows and outflows in order to achieve a stabilizing stock (see Appendix II for the playground scenario).

CO_2 emissions are caused by the burning of fossil fuels and lead to an increase of atmospheric CO_2 concentration. CO_2 absorptions are caused by woods and oceans and decrease atmospheric CO_2 concentration. The Figure below depicts CO_2 emission and CO_2 absorption trajectories between 2010 and 2050.



1. How does CO_2 emission relate to CO_2 absorption between 2010 and 2050 in the Figure above?

□ CO_2 emission is greater than CO_2 absorption.
□ CO_2 emission is smaller than CO_2 absorption.
□ CO_2 emission and CO_2 absorption are equivalent.

2. If CO_2 emission and CO_2 absorption relate to each other as depicted in the Figure above: What happens to atmospheric CO_2 concentration?

□ CO_2 concentration will rise.
□ CO_2 concentration will fall.
□ CO_2 concentration will remain constant.

3. Assuming that the atmospheric CO_2 concentration will fall: What would the corresponding CO_2 emission and absorption trajectories have to look like?

□ CO_2 emission would have to be greater than CO_2 absorption.
□ CO_2 emission would have to be smaller than CO_2 absorption.
□ CO_2 emission would have to be equal to CO_2 absorption.

4. Please sketch your answer to question 3. into the figure below. Draw one line for CO_2 emission and another line for CO_2 absorption trajectories and label them.

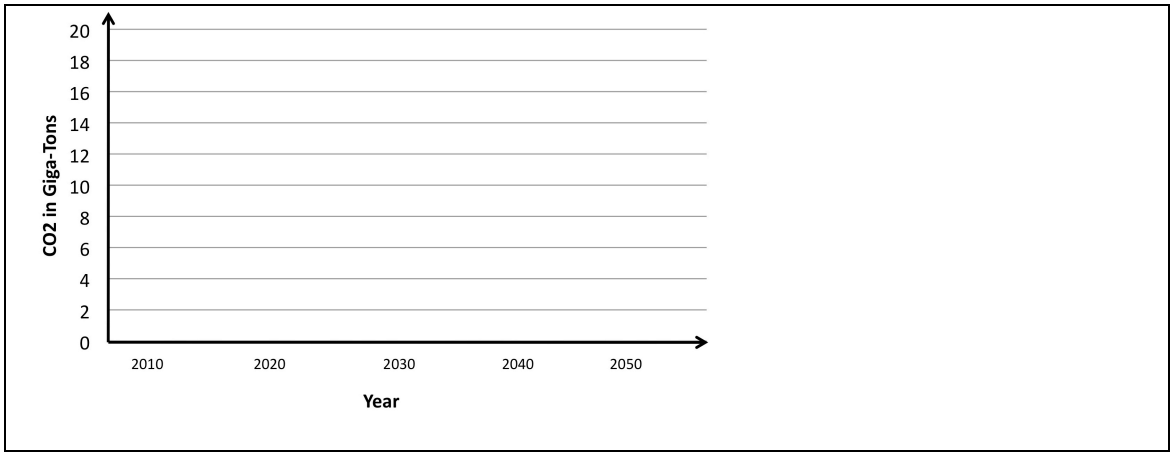


Figure 3. Example of the I/P task: atmospheric CO_2 scenario (translated). Participants are presented with emissions and absorptions trajectories (above). The following sub-tasks test participants' interpretation of graphs (question 1), verbal inferences about the resulting stock (question 2), verbal inferences about emissions and absorptions given a decreasing stock (question 3), and graphical production of the answer to question 3 into the coordinate system (below).

(3)Verbal Tasks: Three SF tasks using three different scenarios (money inside a piggy bank, water inside a bathtub, atmospheric CO_2 concentration) were administered verbally and as a multiple-choice test in order to minimize problems with the task format. In each task, participants first received a short introduction to the problem describing the respective in- an outflows. In the bathtub (piggy bank) task, participants were asked to name the correct strategy in order to achieve a stabilizing (rising) stock, in the CO_2 task, participants needed to determine how the stock reacts if emissions were reduced by a certain amount. We will illustrate the tasks using the bathtub scenario. Participants were shown a picture of a bathtub with a faucet and a drain and were given the following information and questions: *Here you see a bathtub. Water runs into this bathtub through the tap. Meanwhile, water runs out of the bathtub through the drain because it does not seal properly. Imagine, ten minutes ago, you started letting water run into the bathtub and you are now satisfied with the water level. What do you have to do in order to keep the current water level constant?*

a) *You have to open the water tap a little further.*
b) *You have to leave the tap as it is.*
c) *You have to close the water tap a littler further.*

Thus, the bathtub scenario was a verbal translation of the original paradigm comprising in- and outflows and a to-be-stabilized stock. (In all three scenarios, the first three questions asked how the stock will react if inflows are higher, lower, or equal to outflows, for example: *Imagine more water runs into the bathtub than out of the bathtub. How would the water level react?* However, we limit our analyses to the more difficult questions described here since the first three were answered by nearly all participants.) In contrast to the bathtub and piggy bank scenarios, the notation of the CO_2 scenario was slightly scientific, comprising percentage values and some technical information on CO_2 emission and absorption in order to increase the generalizability of results obtained in the verbal tasks (see Appendix III for the piggy bank and CO_2 scenario).

Procedure. Participants were randomly assigned to the graphical (n = 140) or the textual version (n = 30) of the standard tasks. In the graphical version, each participant completed a set of 2 or 3 tasks with different scenarios as part of a larger study that was not connected to SF tasks. Participants were randomly assigned to one set of tasks with presentation order randomized across participants. In the textual version, each participant was randomly assigned to one condition (IS or ~IS). I/P tasks and verbal tasks were completed by n = 107 participants, also as part of other unrelated tasks. Each participant completed both I/P tasks (playground, CO_2) and one randomly assigned verbal task (bathtub, piggy bank, CO_2). Presentation order was randomized across participants both between I/P tasks the and verbal task and within I/P tasks.

Results

For standard tasks, participants' sketched solutions were rated qualitatively correct when they were consistent with SF principles, but they did not have to be quantitatively correct (Cronin, Gonzalez, Sterman, 2009). That is, sketches were rated correct, when, from any point in time onwards, the drawn emission trajectory was higher than (lower than, identical to) the drawn absorption trajectory when the stock was increasing (decreasing, stabilizing). Sketched solutions to question 4 of the I/P tasks were rated correct when they corresponded to the respective answer given to question 3.

For all tasks, presentation order was found to have no impact on solution rates. Furthermore, contrary to our expectation, PM did not have a significant effect on participants´ average correct solutions (M = 15.5% in the ~PM and M = 17.2 % in the PM condition). Thus, solutions were pooled over presentation orders and over both PM conditions, resulting in n = 116, n = 94, n = 33, and n = 47 for the ~IS,~2L; IS,~2L; ~IS, 2L; IS, 2L task, respectively.

To investigate whether and to what extent solutions in the standard task depended on surface features of the task format, a mixed-effects logistic regression model was calculated over all tasks with IS and 2L as fixed factors, participants as random factors, and solution as the dependent variable. A mixed-effects regression model was chosen over regular regression since in our within-subjects design, independence of observations could not be assumed (Baayen, Davidson & Bates, 2008). In line with our expectations, both IS and 2L yielded a significant effect on participants' solutions (Table 1).

Table 1.

Fixed effects coefficients of the logistic mixed effects model for the standard tasks.

Fixed effect	Coefficient	St. error	Lower 95% CI	Upper 95% CI	z-value	p-value
IS	1.886	0.875	0.136	3.636	2.156	0.031
2L	1.224	0.497	0.230	2.218	2.463	0.014

Note. IS = Initial stock given or not; 2L = In- and outflow depicted as two lines vs. as line and dot.

Specifically, solution rates increased from 16% in the original task when neither IS nor

2L were given, to 40% when both IS and 2L were given (solution rates were 29% and 30% for the IS, ~2L and ~IS, 2L task, respectively). The effect of the IS was particularly strong in the textual condition, yielding an average correct solution of M = 80% in the IS, compared to M = 40% in the ~IS condition, χ^2 (1, N = 30) = 5.00, p = .025, indicating that for both graphical and textual task displays, solution rates increased when additional numerical surface information was provided, see Figure 4.

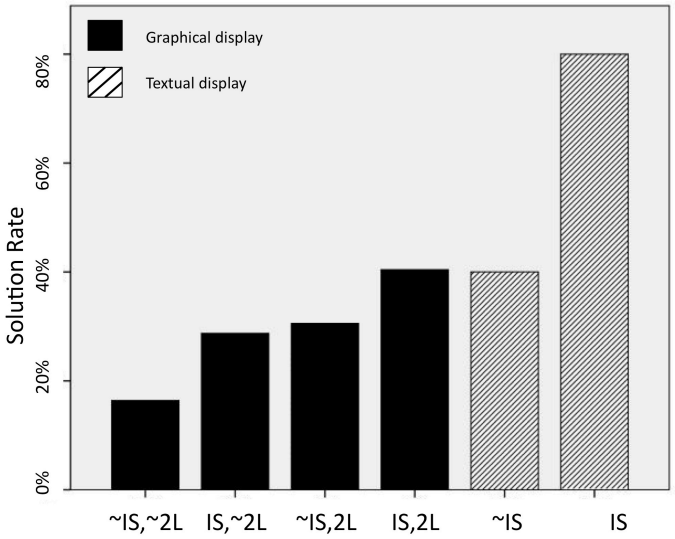


Figure 4. Solution rates for different conditions of both the graphical display (left) and the textual display (right) in the standard task. IS and ~IS denote whether the initial stock was given in the task display, or not. 2L and ~2L denote whether both the in- and outflow trajectory were depicted as singe lines in the task display, or whether only the inflow was depicted as a line, and the outflow was depicted as a single dot. Note that the ~IS, ~2L condition is equivalent to the original task used in previous research (Sterman & Booth Sweeney, 2002, 2007).

In the I/P tasks, the majority of our sample (M = 97%) was able to correctly read and interpret the graphs presented to them (question 1, see Fig. 3). Verbal production tasks were also correctly answered by the majority of participants (M = 83%, M = 89%, for question 2 and 3, respectively). However, in line with our expectations, translating verbal answers into a graphical presentation (question 4) was only accomplished by 57% of the sample. A McNemar test yielded a significant difference between the mean solution rate of both verbal production tasks and the graphical production task, χ^2 (1, N

= 107) = 8.65, p = .003, indicating that for most participants, answers were easier to provide in a verbal than in a graphical format. Unexpectedly, while there were no significant differences between both graphical production tasks (CO_2 vs. playground) for Questions 1-3 ($p > .05$), a McNemar test yielded a significant difference between solution rates of both scenarios in the graphical production task (question 4), χ^2 (1, N = 107) = 16.80, $p < .001$: While 79.3% of the participants were able to sketch their answer in the CO_2 scenario, only 35.4% were able to sketch their answers in the playground scenario. That is, participants were more correct drawing the relation "outflow must be smaller than the inflow" than drawing the relation "outflow must equal inflow". We found a typical mistake in sketching the latter: Instead of drawing two identical lines, 22% of participants drew two parallel lines, resulting in different y-values for in- and outflow. (Note that lines were only rated as parallel, and not as identical if they were at least 0.2 inch apart.)

In line with our hypothesis, the majority of our sample was able to answer SF questions in the verbal task format, yielding an average correct solution of M = 86%. Specifically, solution rates ranged from 98% and 90% (bathtub and piggy bank task, respectively) to 70% (CO_2 task). Thus, SF failure could be reduced when a task format without coordinate systems and graphs and without a focus on quantitative information was used.

Discussion

The present experiment tested whether SF failure can be at least partly explained by specifics of the task formats used in previous research (e.g., Sterman & Booth Sweeney, 2007). Specifically, two caveats were made out.

First, in line with our expectations, both graphical and textual versions of the original paradigm seem to encourage participants to focus on numerical surface features of the task (local search) instead of trying to detect the problem structure (global search): Solution rates varied as a function of the surface information even when the structure remained constant. This was true even for the textual task design, given that solution rates doubled to 80% in the IS condition. Importantly, we do not believe that participants in the IS compared to the ~IS condition were able to detect the problem structure, i.e., we argue that high solution rates in the original paradigm do not reflect

participants' ability to understand the SF structure, just like we argue that low solution rates in the original paradigm (Cronin, Gonzalez & Sterman, 2009) do not reflect participants' inability to understand SF structures. Much rather, we interpret this result as showing that solutions in the original paradigm are highly dependent on surface features of the task format, thus not mirroring participants' actual (mis)understanding of the underlying SF structure but rather their (in)ability to use the surface features of the task, such as being able to sketch their answers or to calculate exact solutions.

In contrast to our expectation, however, solution rates in the original paradigm did not differ as a function of the suggestion of pattern matching in the task design. This result suggests that, at least for this task format using graphs, the use of a pattern matching seems to be an inherent and generic strategy that participants adhere to independent of the task format.

Second, in line with our hypothesis, I/P tasks revealed that the production of graphs in the standard task may have artificially decreased solution rates. We found that solutions to one and the same task were reduced by up to 50% when a graphical compared to a verbal answer was needed. Thus, submitting answers graphically results in a dramatic underestimation of participants' true problem solving abilities.

One task with a stabilizing stock was particularly revealing: In the verbal condition, most participants arrived at the correct solution (inflow equaling outflow); when asked to draw this exact answer into a coordinate system, however, nearly one quarter of our participants sketched two parallel lines. Interestingly, this misconception in the construction of graphs may be able to partially explain the typical mistake in the standard task with stabilizing stock (e.g., Sterman & Booth Sweeney, 2007): Our results suggest that at least some participants may well have the correct verbal representation of the inflow needing to equal the outflow, but then submit a wrong answer by sketching the inflow paralleling the outflow. Thus, the original paradigm using coordinate systems and graphs seems to underestimate participants' ability to grasp SF problems because a potentially error-prone layer is added between participants' mental representations and their submitted answers. On a more general level, this result suggests that, while it may well be reasonable to convey complex information in the form of suitable graphs, participants' understanding of a problem should not be retrieved graphically.

When both caveats (focus on quantitative information, use of coordinate systems) were avoided in the verbal tasks, a vast majority of participants arrived at the correct solution to different SF problems. This result suggests that participants are able to get the qualitative gist of SF problems when they are presented verbally.

Moreover, even the use of the pattern matching heuristic was significantly reduced in the verbal CO_2-task given that 70% of participants correctly answered that the stock increases, even if CO_2 emissions are reduced. In other contexts it was repeatedly shown that participants are able to overcome simple heuristics with practice or insight and, if possible, prefer to make use of the causal structure underlying the problem (Brehmer, 1976; Garcia-Retamero, Wallin & Dieckmann, 2007; Gonzalez, 2004). Similarly, it was assumed before that participants might *either* use the pattern matching heuristic *or* learn to make use of the deep structure of the problem (Cronin, Gonzalez & Sterman, 2009). Our results confirm this hypothesis and generalize previous research on heuristics versus causal structures to the SF context: If SF tasks are presented in such a way that participants have problems understanding their causal structure (standard tasks), they make use of the simple pattern matching heuristic. If, however, tasks are presented in such a way that participants can detect their causal structure (verbal tasks), participants are able to overcome heuristic solutions and arrive at more complex inferences using the SF structure instead.

Limitations

It could be argued that the validity of the verbal tasks is to be doubted, specifically, that they were too simple. Of course, given the high solution rates, verbal tasks were indeed simpler than tasks in the original paradigm. In order to determine whether SF tasks were too simple, however, one needs to determine a) what any SF task needs to assess in order to be valid and b) what might have caused the simplicity of verbal tasks.

Concerning a), SF tasks necessarily need to test whether participants understand SF problems. In our opinion, this means testing whether participants are able to detect the underlying problem structure—"that the quantity of any stock, such as the level of water in a tub, rises (falls) when the inflow exceeds (is less than) the outflow" (Sterman, 2010, p.3)—and whether participants are capable of using their potential understanding of the problem structure, e.g. by making inferences about a stock. Furthermore, such a

test should limit potential misunderstanding of the task format, and solutions should be as independent of specific features of the task format and as close to participants' mental representations as possible. We argue that all those requirements can be met by a qualitative SF task, especially because our results suggested that solutions in the original (quantitative) paradigm are dependent on and participants have problems with certain features of the task format.

Concerning b), higher solution rates of the verbal tasks could have two reasons. First, verbal tasks might have been structurally simpler than SF tasks in the original paradigm. This is not the case, though, since verbal tasks were designed to be structurally equivalent to SF tasks used previously, comprising one stock, one inflow and one outflow with participants determining one or two aspects given the other. That is, verbal tasks were isomorphic to the tasks used in the original paradigm. The second reason could be that, albeit structurally equivalent, the problem structure was given away to the participants. Verbal tasks, however, differed in the extent to which the structure was made explicit to the participant in the answer options, yet solution rates were high even in the most difficult task: Whereas in the piggy bank scenario, the magnitude of the inflow was explicitly related to the magnitude of the outflow, in the bathtub scenario, only the inflow was mentioned in the answer options, and participants needed to establish the relation between in- and outflow on their own. Moreover, in the CO_2 scenario, this relation even needed to be established for a specific amount of inflow reduction. Thus, at least for the bathtub and CO_2 scenario, participants themselves needed to bring together the decisive aspects of the SF structure rendering these tasks a direct translation of the original paradigm. Consequently, we argue that the validity of the verbal tasks is diminished neither because of their qualitative format, nor because they were structurally simpler nor because they gave away the problem structure. We reason that verbal tasks are easier because their task format makes it easier for participants to detect the underlying SF structure—a structure which, in turn, is actually quite simple.

Implications

Present findings have various implications for the communication of SF problems such as the accumulation of debts, or the accumulation of CO_2 in the

atmosphere. Given the nature of our sample, members of the general public (and not only highly educated students) are able to understand SF problems when presented qualitatively. Based on these findings, we suggest that display formats used in media reports should be modified. Specifically, we suggest that quantitative information is reduced to a minimum in order to render abstract topics more accessible and to communicate risk more effectively. This is even more important given that the way information is presented does not only affect understanding of the problem, but also the quality of decision-making (Covey, 2011). For example, it was reasoned that people's misunderstanding of SF structures inherent to climate change could explain their lack of motivation to contribute to climate change mitigation (Sterman, 2008). It is up to future research to determine, however, whether increasing people's understanding of climate change by presenting its SF structures verbally, will also lead to an increased motivation to contribute to its solution.

Conclusion

Based on the presented findings, the previously well-established SF failure seems to be attributable to the salience of quantitative information in the information display on the one hand and on participants' difficulties to construct line graphs on the other. Consequently, in the original paradigm, participants were not able to demonstrate their verbal understanding of SF problems. If SF problems are considered as understood only when participants uncover the SF structure from quantitative data (be they presented textual or graphical), one can say that people have severe difficulties understanding SF problems. However, if SF problems are considered as understood also when participants uncover the SF structure from qualitative data, our results suggest that people do understand SF problems. That is, SF failure is valid for, but confined to, a specific task type, namely one that conceives of the SF structure as quantitative rather than qualitative. On a more general level, our findings contribute to a long history of psychological research showing that most people are able to effectively deal with even highly complex tasks, as long as they are presented in an intuitive and accessible format.

References

Bassok, M. 2001. Semantic alignments in mathematical word problems. In Gentner, D., Holyoak, K. & Kokinov, B.N. (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 401–433). Cambridge, MA: MIT Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* **59**: 390-412.

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

Brehmer, B. 1976. Learning complex rules in probabilistic inference tasks. *Scandinavian Journal of Psycholog,* **17**: 309–312.

Carpenter, P. A., & Shah, P. 1998. A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied* **4**: 75-100.

Cheng, P. W., & Holyoak, K. J. 1985. Pragmatic reasoning schemas. *Cognitive Psychology* **17**: 391-416.

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

Covey, J. 2011. The effects of absolute risks, relative risks, frequencies, and probabilities on decision quality. *Journal of Health Communication* **16**: 788-801.

Fischer, M. H. 2000. Do irrelevant depth cues affect the comprehension of bar graphs? *Applied Cognitive Psychology* **14**: 151-162.

Garcia-Retamero, R., Wallin, A. & Dieckmann, A. 2007. Does causal knowledge help us be faster and more frugal in our decisions? *Memory & Cognition* **35***:* 1399-1409.

Gattis, M., & Holyoak, K. J. 1996. Mapping conceptual to spatial relations in visual reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **22***:* 231-239.

Gigerenzer, G., & Hoffrage, U. 1995. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review* **102***:* 684-704.

Gonzalez, C. 2004. Learning to make decisions in dynamic environments: Effects of time constraints and cognitive abilities. *Human Factors* **46**: 449-460.

Guthrie, J. T., Weber, S., & Kimmerly, N. 1993. Searching documents: Cognitive processes and deficits in understanding graphs, tables, and illustrations. *Contemporary Educational Psychology* **18**: 186-221.

Hegarty, M., Canham, M. S., & Fabrikant, S. I. 2010. Thinking about the weather: How display salience and knowledge affect performance in a graphic inference task. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **36**: 37-53.

Hoffrage, U., & Gigerenzer, G. 1998. Using natural frequencies to improve diagnostic inferences. *Academic Medicine* **73**: 538-540.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. 2000. Communicating statistical information. *Science* **290***:* 2261-2262.

Kotovsky, K. K., Hayes, J. R., & Simon, H. A. 1985. Why are some problems hard? Evidence from Tower of Hanoi. *Cognitive Psychology* **17**: 248-294.

Landy, D., & Goldstone, R. L. 2007. How abstract is symbolic thought? *Journal of Experimental Psychology: Learning, Memory, and Cognition* **33**: 720-733.

Novick, L. R., & Catley, K. M. 2007. Understanding phylogenies in biology: The influence of a Gestalt perceptual principle. *Journal of Experimental Psychology: Applied* **13**:197-223.

O'Brien, D. P., Noveck, I. A., Davidson, G. M., & Fisch, S. M. 1990. Sources of difficulty in deductive reasoning: The THOG task. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology* **42**: 329-351.

Shah, P., & Carpenter, P. A. 1995. Conceptual limitations in comprehending line graphs. *Journal of Experimental Psychology: General* **124**: 43-61.

Shah, P., & Hoeffner, J. 2002. Review of graph comprehension research: Implications for instruction. *Educational Psychology Review* **14**: 47-69.

Sterman, J.D. & Booth Sweeney, L. 2002. Cloudy Skies: Assessing public understanding of global warming. *System Dynamics Review* **18**: 207-240.

Sterman, J.D. & Booth Sweeney, L. 2007. Understanding public complacency about climate change: adults' mental models of climate change violate conservation of matter. *Climatic Change* **80**: 213-238.

Sterman, J.D. 2008. Risk communication on climate: Mental models and mass balance. *Science* **23**: 532-533.

Sterman, J.D. 2010. Does formal system dynamics training improve people's understanding of accumulation? *System Dynamics Review* **26**: 313-334.

Sterman, J.D. 2011. Communicating climate change risks in a skeptical world. *Climatic Change* **108**: 811–826.

Wainer, H. 1992. Understanding graphs and tables. *Educational Researcher* **21**: 14-23.

Winn, W. 1991. Learning from maps and diagrams. *Educational Psychology Review* **3**: 211-247.

Zhang, J., & Norman, D. A. 1994. Representations in distributed cognitive tasks. *Cognitive Science* **18**: 87-122.

Zhang, J., & Norman, D. A. 1995. A representational analysis of numeration systems. *Cognition* **57**: 271-295.

Appendix I: Scenarios used for the standard tasks (translated).

Department store scenario

The following Figure depicts how many people are inside a department store over the course of 10 minutes.

The following Figure depicts the amount of people entering and leaving the above department store. Please draw both how many people must enter and how many people must leave the department store from minute 5 to 10 in order to achieve the above stock (several solutions are possible, please draw only one).

IS condition: At the beginning, 32 people are inside the store.

Members of a club scenario

The following Figure depicts the amount of members of a club over the course of 10 years.

The following Figure depicts the amount of people joining and leaving the club. Please draw both how many people must join the club and how many people must leave the club from year 5 to 10 in order to achieve the above stock (several solutions are possible, please draw only one).

Guests on a party scenario

The following Figure depicts the amount of guests on a party over the course of 10 minutes.

The following Figure depicts the amount of people coming to and leaving the party. Please draw both how many people must come to and how many people must leave the party from minute 5 to 10 in order to achieve the above stock (several solutions are possible, please draw only one).

Nuts collected by a squirrel scenario

The following Figure depicts the amount of nuts of a squirrel over the course of 10 days.

The following Figure depicts how many nuts the squirrel collects and eats. Please draw both how many nuts the squirrel must collect and how many nuts the squirrel must eat

from day 5 to 10 in order to achieve the above stock (several solutions are possible, please draw only one).

IS condition: At the beginning, the squirrel has 3 nuts.

Appendix II: Playground scenario used for the I/P tasks (translated).

The following Figure depicts the amount of children entering and leaving a playground.
1. How does the amount of children entering the playground relate to the amount of children leaving the playground?

[] The amount of children entering equals the amount of children leaving the playground.

[] The amount of children entering is higher than the amount of children leaving the playground.

[] The amount of children entering is lower than the amount of children leaving the playground.

2. If the amount of children entering and leaving the playground relate to each other as depicted above: How will the amount of children who actually are on the playground develop over time?

[] The amount of children on the playground will rise.

[] The amount of children on the playground will fall.

[] The amount of children on the playground will remain constant.

3. Assuming that the amount of children on the playground will remain constant: How would the amount of children entering have to relate to the amount of children leaving?

[] The amount of children entering would have to be greater than the amount of children leaving.

[] The amount of children entering would have to be equal to the amount of children leaving.

[] The amount of children entering would have to be less than the amount of children leaving.

3. Please sketch your answer to question 3 into the Figure below. (Several solutions are possible, please sketch only one).

Appendix III: Scenarios used for the verbal tasks (translated)

Piggy Bank Scenario

Image you have a piggy bank. Each month, you throw money into the piggy bank, and you also take some money out of the piggy bank. Imagine, you want to buy yourself a book worth 20 Euro. You count the money inside your piggy bank and notice that you currently have 10 Euro. What do you need to do to ensure the amount of money will increase to 20 Euro?

a) You have to take less money out of the piggy bank than you throw into it.
b) You have to take more money out of the piggy bank than you throw into it.
c) You have to take out as much money as you throw into the piggy bank.

CO_2 Scenario

CO_2 emissions are caused by the burning of fossil fuels and lead to an increase of atmospheric CO_2 concentration. CO_2 absorptions are caused by woods and oceans and decrease atmospheric CO_2 concentration. CO_2 emissions are currently twice as high as CO_2 absorptions. Imagine, emissions were reduced by 30%: How would the atmospheric CO_2 concentration react?

a) Atmospheric CO_2 concentration would increase
b) Atmospheric CO_2 concentration would decrease
c) Atmospheric CO_2 concentration would remain constant