System Dynamics Review
System Dynamics Review (2011)
Published online in Wiley Online Library
(wileyonlinelibrary.com) **DOI**: 10.1002/sdr.470

Understanding stocks and flows through analogy

Cleotilde Gonzalez* and Hau-yu Wong

Abstract

Although it has been suggested that people use the wrong cognitive procedures in solving stock and flow (SF) problems, we know little of what these mental procedures are. We present two experiments aimed at demonstrating the influence of analogical reasoning on SF failure. Results of Experiment 1 show that SF failure decreases when people are asked to compare problems that share behavioral similarity (common relations). However, the benefit of behavioral similarity depends on the surface similarity (common superficial object attributes). Results from Experiment 2 demonstrate that when the behavioral characteristics of the problems are unknown by the participants, the process of comparing two problems with behavioral similarity improves responses to a subsequent SF problem, regardless of the surface similarity between the problems. Surface similarity helps only when the behavioral similarity between the problems is already known. Implications for training and education in system dynamics are discussed. Copyright © 2011 System Dynamics Society

Syst. Dyn. Rev. (2011)

Introduction

Our understanding of the mental procedures used while solving accumulation (stock) and rates of change (flow) problems is fundamental to a better appreciation of decision making in dynamic environments. For example, a manager must understand how inventory accumulates during production and decreases during shipments in order to make optimal production and marketing decisions; the federal government must interpret the national debt as the federal deficit accumulates in order to release effective economic policies; a person working on maintaining a healthy weight needs to understand how calorie consumption accumulates and how it decreases through exercise, and many other examples. Thus understanding and managing accumulation is an important concept at many levels of human life: the individual, organizational, and social.

Unfortunately, there is increasing and robust evidence of a fundamental lack in the human understanding of accumulation and rates of change: a difficulty called the *stock–flow (SF) failure* (Cronin *et al.*, 2009). The SF failure occurs even in simple problems, such as evaluating the level of water in a bathtub given the amounts flowing in (inflow) and out (outflow) over time (Booth Sweeney and Sterman, 2000). Researchers have used simple problems to ask individuals for their basic interpretations of a stock's behavior. For example, researchers often use graphical representations of the inflow and outflow over time, and ask students to answer questions about the stock or to draw it. Despite the simplicity of these problems, individuals with strong mathematical backgrounds exhibit poor performance: less than 50 percent of them answer the stock questions correctly (Cronin and Gonzalez, 2007).

* Correspondence to: Cleotilde Gonzalez, Department of Social and Decision Sciences, Carnegie Mellon University, 208 Porter Hall, 5000 Forbes Avenue, Pittsburgh, PA 15213, U.S.A. E-mail: coty@cmu.edu

Received 28 April 2010; *Accepted 21 June 2011*

It has been suggested that people use the "wrong representation" to think about SF (Cronin *et al*., 2009), and students often draw a stock that replicates the pattern of inflow or net flow, while ignoring the effect that both inflow and outflow would have on the accumulation over time. These mental procedures that lead to erroneously assuming that the stock behaves like the flows was termed the "correlation heuristic" (Cronin *et al*., 2009). Although the correlation heuristic seems to be robust in SF problems (Booth Sweeney and Sterman, 2000; Cronin and Gonzalez, 2007; Cronin *et al*., 2009), we know little about the mental procedures people use in solving these problems and why.

Cronin *et al*. (2009) suggested that early education may reinforce the impression that the relations between the flows and stock are proportional (i.e., they correlate positively), when in fact they are nonlinear. Although the knowledge needed to answer SF problems correctly is basic,[1] the education literature has documented a "linearity" illusion, which refers to the extensive attention paid to linear reasoning in mathematical education and the encouragement to apply linear models even in non-applicable situations (Van Dooren *et al*., 2003, 2004). Errors related to correlational reasoning in problem solving are widespread in the education literature (Harel *et al*., 1992; Ben-Zeev and Star, 2001; Van Dooren *et al*., 2005). Thus it is possible that individuals attempting to solve SF graphical problems are responding in ways that have been ingrained due to early education's linearity illusion and correlational reasoning. Analogies are typically drawn from a well-understood situation to a poorly understood one (Kurtz *et al*., 2001), and it is possible that individuals see some similarities between SF problems and their problems in early education, and attempt to draw an analogy between them.

The analogical reasoning literature usually distinguishes between problems that are *superficially* similar and those that share *behavioral* similarity (Holyoak and Koh, 1987; Medin *et al*., 1995; Lowenstein *et al*., 1999; Thompson *et al*., 2000; Kurtz *et al*., 2001).[2] The distinction between surface and behavioral similarity is usually based on the relevance of various features in a problem to its goal attainment. Surface dissimilarity between two problems involves differences in attributes that do not influence goal attainment, while behavioral dissimilarity indicates the presence of differences in the causal relations within a problem (Holyoak and Koh, 1987). That is, surface similarity is based on the mere appearance between two objects, whereas behavioral similarity is based on the function, matching relations, and final goal of the problems even when they do not appear to be similar.

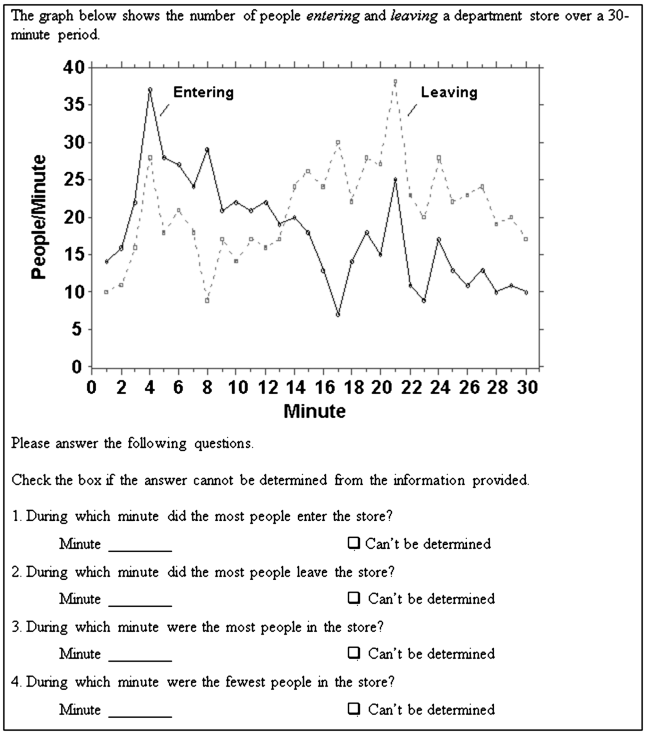
A person's ability to differentiate between a problem's surface features and its behavioral similarity is often lacking. Problem solvers often mistake surface features as functionally important to formulating the solution to a problem (Medin *et al*., 1995; Lowenstein *et al*., 1999; Thompson *et al*., 2000; Kurtz *et al*., 2001). On the other hand, studies of expertise in domains such as physics have demonstrated a shift from representations based on surface features in those less experienced in a task to representations based on behavioral features for those more experienced (e.g., Chi *et al*., 1981). These studies suggest that those with more experience should be able to use analogy more successfully and determine the behavioral similarities in a task, rather than being distracted by the surface similarities (Holyoak and Koh, 1987).

A goal of this paper is to test the use of analogies in solving SF problems. Based on findings from the analogy literature, we expect that participants in SF studies may be using the surface similarity between SF problems and the linear reasoning problems in their mathematical education, and may not be able to see that the behavioral characteristics of linear and nonlinear problems are quite different. We aim at determining how participants use their experience with similar problems (both in surface and behavioral similarity) to solve a subsequent SF problem.

The department store task

The "department store" (DS) task, illustrated in Figure 1, was initially reported by Sterman (2002), and it has been used to demonstrate the SF failure and correlation heuristic in recent research (Cronin and Gonzalez, 2007; Cronin *et al.*, 2009). This task presents participants with a graph showing the number of people entering and leaving a department store each minute over a 30-minute interval. The system involves a single stock (the number of people in the store) with an inflow (people entering) and outflow (people exiting). There are no feedbacks delays, no stochastic events, or any other elements of dynamic complexity that proved difficult in prior research (Sterman, 2002).

Participants are asked four questions (Figure 1). The first two questions test whether participants can read the graph and correctly distinguish between the inflow and outflow. The third and fourth questions test whether participants can infer the stock's behavior from the behavior of the flows. The SF failure is based on the low percentage of correct



The graph below shows the number of people *entering* and *leaving* a department store over a 30-minute period.

Please answer the following questions.

Check the box if the answer cannot be determined from the information provided.

1. During which minute did the most people enter the store?

Minute ________ □ Can't be determined

2. During which minute did the most people leave the store?

Minute ________ □ Can't be determined

3. During which minute were the most people in the store?

Minute ________ □ Can't be determined

4. During which minute were the fewest people in the store?

Minute ________ □ Can't be determined

Fig. 1. The department store (DS) problem

DOI: 10.1002/sdr

responses to questions 3 and 4, which was about 42 percent in past studies (Sterman, 2002; Cronin and Gonzalez, 2007; Cronin *et al*., 2009).

Flows that follow the characteristics in Figure 1 result in an inverted U-shape stock curve that has a right tail slightly lower than the left tail. Thus this is a nonlinear problem in which the linear or correlation heuristic does not apply. One cannot correctly estimate the stock based on either the path of the inflow (people entering the store) or the path of the outflow (people leaving the store) alone. Past research has demonstrated that people continue to follow the correlation heuristic in this case: a common mistake for question 3 is to answer with the time at which the net inflow was maximum (where the gap between the inflow and outflow is largest; $t=8$), and a common mistake in question 4 is to answer with the time at which the net outflow is maximum ($t=17$).

The next experiments, based on analogy research, will attempt to answer the following questions. Can the comparison of two analogical cases help participants see the behavioral similarity of a subsequent DS task and thus answer questions 3 and 4 correctly? What type of similarities in the analogous problems (behavioral or surface) would help participants to achieve the best performance in a subsequent DS task?

Experiment 1: Helping to see behavioral similarity through analogy

We expect that analogical comparison of problems with behavioral similarity to the DS problem can help people “see” the behavioral similarity and improve the overall percentage of correct responses to questions 3 and 4 in a subsequent DS task.

Based on the analogy literature, we define two SF problems as behaviorally similar if they share the same relations of inflows and outflows to result in a similar stock shape over time. In the case of the DS task, a problem would be behaviorally similar to another one if it contains the same relationships of inflow and outflow to produce an inverted U-Shape curve of the stock with the right tail of the U slightly lower than the left tail. A problem would be similar in surface to the DS task if it shares features that are irrelevant to obtain the resulting stock in the DS task. For example, the total time intervals on the *x*-axis might be the same as in the DS task (e.g., 30 minutes) yet produce a very different stock shape; the particular values of the inflow and outflow per unit of time on the *y*-axis might be very similar to those in the DS task, yet the resulting stock shape may be different.

The process of comparing two problems allows people to apply a “mapping” between them that would accent the characteristics shared by both (Gentner and Markman, 1997). The assumption is that comparison of two analogous examples helps people to consider commonalities in the behavioral characteristics of different problems, regardless of surface differences. For example, Thompson *et al*. (2000) employed two negotiation problems during training that shared behavioral similarity but were not similar in surface, in order to promote extraction of only the behavioral relations. Even without instructions on how to make comparisons, participants were three times more likely to employ the strategy taught by the two analogous examples in a final negotiation problem given after the comparison (Thompson *et al*., 2000). In addition, they demonstrated better overall understanding of the underlying behavioral principle, when they were told that the quality of their comparisons would be rated. Thus, in order to understand the behavioral correspondence of any two problems, people do not require explicit explanations of the underlying principle to take full advantage of the comparison process.

DOI: 10.1002/sdr

By using the comparison of two problems with behavioral similarity, we might be highlighting the behavioral characteristics in the DS problem. In past research, many failed to recognize the behavioral similarity of the SF problems even after being directed to keep track of the accumulation of the stock for each period over time (Cronin *et al.*, 2009). The process of comparison proposed by Thompson *et al.* (2000) is expected to help promote spontaneous analogical transfer to a subsequent DS task. Thus, we expect that:

$H_{1.}$ Participants would respond more accurately to accumulation questions in the DS problem after they first compare two SF problems that share behavioral similarity with the DS problem.

Given that the analogy literature suggests that problem solvers often mistake surface features as functionally important to formulating a problem's solution (Kurtz *et al.*, 2001), it is also possible that if people become too focused on surface attributes they would formulate solutions based only on the surface features of the comparison problems (Gentner, 1983; Gentner and Markman, 1997). In general, transfer may be significantly impaired when either a lack of surface or behavioral similarity is provided during the comparison phase (Holyoak and Koh, 1987). When target problems share both surface and behavioral similarity, we expect that a large majority of participants would be able to generate a correct answer in the novel problem. Thus, we expect that:

$H_{2.}$ Participants would respond more accurately to the stock questions in the DS problem after they compare two SF problems that share both surface and behavioral similarity to the DS problem.

Experimental design

The design involved the comparison of two identical problems, whose graphical representations were either similar or not similar in surface and behavioral characteristics to the DS problem. After the comparison, participants were asked to solve the DS problem by answering the four questions (Figure 1). Participants were assigned to one of four experimental conditions (Figure 2): behavioral and surface similarity (B & S); behavioral similarity but not surface similarity (B & NOT S); not behavioral similarity but surface similarity (NOT B & S); and not behavioral similarity and not surface similarity (NOT B & NOT S).

The two identical problems given in the B & S condition were essentially identical to the DS task. The two were behaviorally similar to the DS task (produced the same resulting shape of stock as the DS task), as well as similar in surface to the DS task (it had the same number of time periods and values of inflow and outflow over time). In B & NOT S, the two identical problems given for comparison shared behavioral similarity to the DS task because they produced an inverted U-shaped stock curve, and they were not similar in surface to the DS task because the graphs appeared superficially different from the DS task: they had 14 time periods instead of the original 30 in the DS task, and the values of the inflow and outflow were different. In NOT B & S, the two identical problems given were not similar in behavioral characteristics to the DS task because the resulting stock was a linearly increasing curve rather than an inverted U-shape, and they were similar in surface to the DS task because they appeared to be superficially similar to the DS task: the graphs had the same number of time periods and similar values of inflows and outflows. Finally, in NOT B & NOT S, the two identical problems given were not similar in both

DOI: 10.1002/sdr

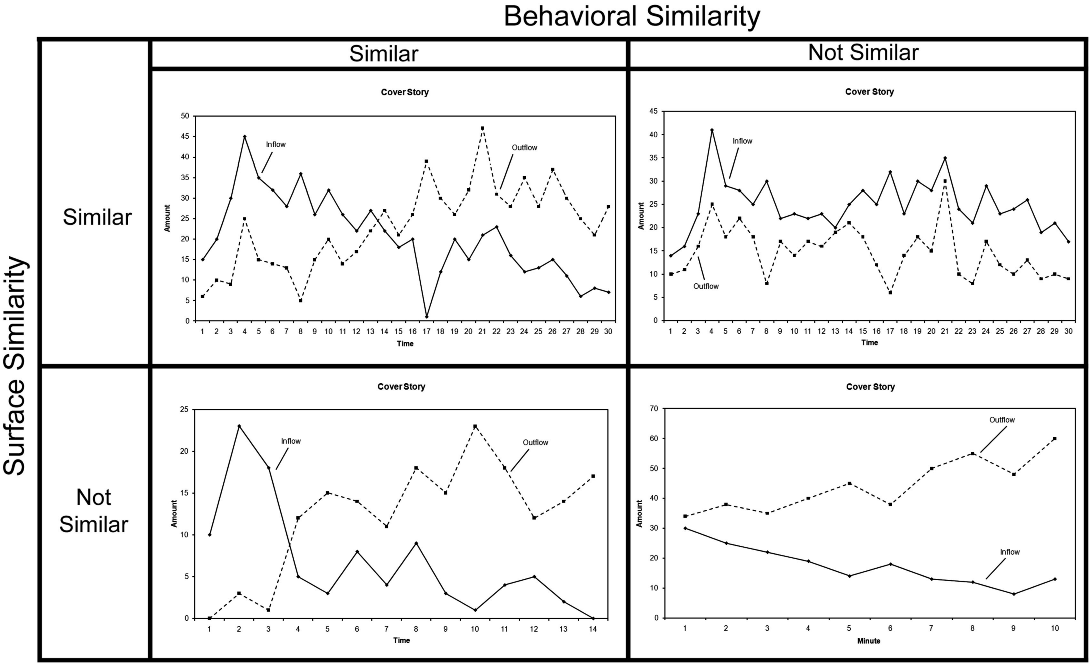


Fig. 2. Experimental design for Experiment 1. Participants were given two identical graphs corresponding to one of the four conditions: similar in behavioral characteristics and similar in surface (B & S); similar in behavioral characteristics and not similar in surface (B & NOT S); not similar in behavioral characteristics and similar in surface (NOT B & S); and not similar in behavioral characteristics and not similar in surface (NOT B & NOT S). The two graphs were given under two different cover stories (the bank account and the water tank)

behavioral and surface characteristics to the DS task: the resulting stock curve is logarithmically decreasing, and the number of time periods and general superficial appearance of the curves is different from the DS problem.

The two comparison problems within each condition were identical to each other, but were presented using two different cover stories: a bank account with money being deposited or withdrawn, and a water tank with water filling or draining away. Providing participants with identical problems but two different cover stories was thought to best facilitate their extraction of the problem's behavioral similarity, as shown in Thompson *et al.* (2000). The main dependent variable was the performance on the four questions in the subsequent DS problem.

Participants

Participants were recruited from the local universities using electronic communication boards and the recruitment website of our laboratory (*N*=63). The average age of this group was 22.59 years (SD=3.57), ranging from 18 to 34 years. 56 percent were males, and 54 percent of all participants majored in mechanical and electrical engineering and other sciences, while the rest were students with non-technical majors. Participants were randomly assigned to one of the four conditions: B & S (*N*=20); B & NOT S (*N*=12); NOT B & S (*N*=18); and NOT B & NOT S (*N*=13).

Procedure

During the comparison phase, participants were given two problems simultaneously, each on a separate page. Each problem included identical sets of instructions and two graphs showing the *Flows* of two identical problems belonging to one of the four conditions. For one example of the full instructions for the B & NOT S condition, see Appendix A. In the instructions, participants were directed, step by step, to locate the maximum inflow and maximum outflow on a graph given to them. Then, they were made aware of the rules of accumulation, highlighting the period of time in which the inflow was greater than the outflow and the period in which the outflow was greater than the inflow. This was important to do so that participants became aware of the shape of the accumulation curve, which they were asked to sketch in a separate blank space. The instructions were identical regardless of the experimental conditions. The correct responses to complete these tasks were provided in the description of each problem. Thus participants only had to read the instructions carefully and follow them, before writing down their comparison of the two problems.

Participants were then asked to write down the similarities between the two source problems using concrete details, and to look for a general principle that would best describe the two problems. They were asked to look at the two problems as they made the comparison, rather than doing so from memory. After participants wrote down the similarities and common features, the experimenter collected these and the comparison sheet, and immediately distributed the DS problem. Participants were allowed to spend an unlimited amount of time on either task, but spent on average about 20 minutes in total.

Results

Comparison phase

We first determined whether participants sketched the correct path of the stock in each of the two comparison problems (the two stocks were identical within each condition). We graded the shape of the stock as right or wrong. It was correct if the participant sketched the right shape of the stock and incorrect if they drew it differently in any way from the correct shape. Using visual inspection of their stock paths, we also coded the erroneous responses to determine whether the path participants drew matched the pattern of the inflow, outflow, or net flow; and whether the stock and inflow or net flow were correlated or not in a similar process to the one in Cronin *et al.* (2009).

Despite the step-by step strategy given to the participants so that all information needed to answer the questions correctly was provided, we found that only 61.9 percent of participants drew the stock's shape correctly in at least one of the comparison problems. The proportion was very similar across the different experimental conditions and demonstrated that the conditions were at about the same level of difficulty (Cronin *et al.*, 2009): B & S (60 percent); B & NOT S (75 percent); NOT B & S (61.1 percent); and NOT B & NOT S (53.8 percent). Most interestingly, the majority (87.5 percent) of those that drew the stock incorrectly on either or both comparison problems followed the correlation heuristic by drawing a stock that matched the pattern of the inflow or the net flow. The percentages of incorrect responses that followed the correlation heuristic per condition were 87.5 percent (B & S); 66.7 percent (B & NOT S); 100 percent (NOT B & S); and 83.3 percent (NOT B & NOT S). Thus most of those that drew the stock incorrectly followed the correlation heuristic, regardless of the condition they were assigned to.

DOI: 10.1002/sdr

The DS problem

Table 1 presents the distribution of responses to the four DS questions for each of the four conditions. The general proportion of correct responses to each question is comparable to that of previous studies (on average, 92.1 percent, 90.5 percent, 54 percent, and 55 percent answered questions 1, 2, 3, and 4 correctly, compared to 95.9 percent, 94.7 percent, 43.9 percent, and 41.2 percent respectively in Cronin *et al*. (2009). Table 1 also shows the different types of errors made in each question. For example, the predominant type of incorrect response given for question 3 was the maximum net inflow (the point of time with the maximum difference between inflow and outflow). A total of 14 incorrect responses to question 3 were of this type. For question 4, the most common error (nine of the incorrect responses) was the maximum net outflow (the point of time where there is the maximum difference between outflow and inflow). These results reaffirm findings about the correlation heuristic found in Cronin *et al*. (2009). The next most frequent type of error made was the "cannot be determined" response to questions 3 and 4.

We conducted a chi-square analysis on the independent effects of surface and behavioral similarity between problems on the percentage of correct responses in the DS problem. We also conducted several chi-square analyses according to participants' success in drawing the stock during the comparison phase (see Table 2).

As shown in Table 2, behavioral similarity of the comparison problems to the DS problem determined the accuracy of responses in the DS problem. Overall, the proportion of correct responses to question 3 was greater (65.6 percent) when participants compared problems behaviorally similar to the DS problem than when they compared problems with no behavioral similarity (41.9 percent), $\chi^2(1)=3.56$, $p<0.05$. Similarly, the proportion of correct responses to question 4 was greater (71.9 percent) when participants compared problems behaviorally similar to the DS problem than when they compared problems with no behavioral similarity (38.7 percent), $\chi^2(1)=7.01$, $p<0.01$. Thus Hypothesis 1 is supported: the comparison of problems that are behaviorally similar to the DS problem resulted in more accurate responses in the DS problem than when they compared problems with no behavioral similarity.

Interestingly, when the responses were separated according to participants' ability to drawthe stocks correctly during comparison, results show that behavioral similarity was significant only in the case of those that drew the stock incorrectly. The proportion of correct responses to question 4 of the DS task for participants that drew the stock incorrectly was significantly greater (72.7 percent) when participants compared problems behaviorally similar to the DS problem than when they compared not behaviorally similar problems (23.1 percent), $\chi^2(1)=5.92$, $p<0.05$. There was no advantage of behavioral similarity in the cases where participants drew the stock shape correctly during the comparison phase.

As Table 2 also shows, surface similarity between the comparison and DS problems did not influence the proportion of correct responses to the subsequent DS task.

Analyses of the interaction effects of both surface and behavioral similarity (Table 3) indicate the advantage of behavioral similarity to answering questions 3 and 4 in the DS problem: behavioral similarity is only relevant when the comparison problems are also similar in surface. The proportion of correct responses to question 3 was greater (75 percent) when participants compared problems that were similar, in both behavior and surface, to the subsequent DS problem than when they compared problems that were not similar in both (38.9 percent), $\chi^2(1)=5.07$, $p<0.05$. Also, the proportion of correct

Table 1. Distributions of responses to the four questions in the DS problem after the comparison in each of the four conditions in Experiment 1. The top part of the table shows the proportions of correct responses in each question and condition; the shaded cells contain the frequencies of correct answers to each of the four questions. The bottom part of the table shows the proportion of errors of different types

	Q1: Most entering?				Q2: Most leaving?				Q3: Most in store?				Q4: Fewest in store?			
	B & S	B & NOT S	NOT B & S	NOT B & NOT S	B & S	B & NOT S	NOT B & S	NOT B & NOT S	B & S	B & NOT S	NOT B & S	NOT B & NOT S	B & S	B & NOT S	NOT B & S	NOT B & NOT S
Correct responses																
Q1:Most entering: *t*=4	19 95%	12 100%	17 94.4%	10 76.9%	0	0	0	0	0	0	1 5.6%	0	0	0	0	0
Q2: Most leaving: *t*=21	0	0	0	1 7.7%	19 95%	12 100%	15 83.2%	11 84.6%	0	0	0	0	0	0	0	0
Q3: Most in store: *t*=13	0	0	0	0	0	0	0	0	15 75%	6 50%	7 38.9%	6 46.2%	0	0	3 16.7%	1 7.6%
Q4: Fewest in store: *t*=30	0	0	0	0	0	0	0	0	0	0	2 11.1%	0	16 80%	7 58.3%	6 33.3%	6 46.2%
Errors																
Max. net inflow (entering − leaving): *t*=8	1 5%	0	0	2 15.4%	0	0	0	0	3 15%	6 50%	2 11.1%	3 23.1%	0	0	1 5.5%	0
Max net outflow (leaving − entering): *t*=17	0	0	0	0	1 5%	0	1 5.6%	2 15.4%	0	0	2 11.1%	1 7.6%	1 5%	5 42.7%	2 11.1%	2 15.4%
Initial in store *t* =1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3 16.7%	0
Can't be determined	0	0	1 4.6%	0	0	0	1 5.6%	0	2 10%	0	4 22.2%	3 23.1%	3 15%	0	3 16.7%	4 30.8%
Other	0	0	0	0	0	0	1 5.6%	0	0	0	0	0	0	0	0	0
Total	20	12	18	13	20	12	18	13	20	12	18	13	20	12	18	13

Table 2. Experiment 1. Analyses of the effects of SURFACE and BEHAVIORAL similarities on the percentage of correct responses to the DS problem. Statistical differences between correct and incorrect responses are in bold

		Surface			Behavioral		
		N=63			N=63		
	Questions	Similar	Not similar	$\chi^2(1)$	Similar	Not similar	$\chi^2(1)$
Overall	Q1: Highest inflow?	94% (36 of 38)	88% (22 of 25)	0.94	96.9% (31 of 32)	87.1% (27 of 31)	2.06
	Q2: Highest outflow?	89.5% (34 of 38)	92% (23 of 25)	0.11	96.9% (31 of 32)	83.9% (26 of 31)	3.09
	Q3: Most in stock?	57.9% (22 of 38)	48% (12 of 25)	0.59	65.6% (21 of 32)	41.9% (13 of 31)	**3.56***
	Q4: Least in stock?	57.9% (22 of 38)	52% (13 of 25)	0.21	71.9% (23 of 32)	38.7% (12 of 31)	**7.01****
Correct stock shape	Q1: Highest inflow?	100% (23 of 23)	93.8% (15 of 16)	1.48	100% (21 of 21)	94.4% (17 of 18)	1.19
	Q2: Highest outflow?	91.3% (21 of 23)	93.8% (15 of 16)	0.08	100% (21 of 21)	83.3% (15 of 18)	3.79
	Q3: Most in stock?	69.6% (16 of 23)	56.2% (9 of 16)	0.73	71.4% (15 of 21)	55.6% (10 of 18)	1.06
	Q4: Least in stock?	65.2% (15 of 23)	56.2% (9 of 16)	0.32	71.4% (15 of 21)	50% (9 of 18)	1.88
Incorrect stock shape	Q1: Highest inflow?	86.7% (13 of 15)	77.8% (7 of 9)	0.32	90.9% (10 of 11)	76.9% (10 of 13)	0.84
	Q2: Highest outflow?	86.7% (13 of 15)	88.9% (8 of 9)	0.03	90.9% (10 of 11)	84.6% (11 of 13)	0.22
	Q3: Most in stock?	40% (6 of 15)	33.3% (3 of 9)	0.11	54.5% (6 of 11)	23.1% (3 of 13)	2.52
	Q4: Least in stock?	46.7% (7 of 15)	44.4% (4 of 9)	0.01	72.7% (8 of 11)	23.1% (3 of 13)	**5.92***

*$p<0.05$; **$p<0.01$.

Table 3. Analysis of the interaction effects of both surface and behavioral similarity

			Behavioral		
		Questions	Similar	Not similar	$\chi^2(1)$
Surface	Similar	Q1: Highest inflow?	95% (19 of 20)	94.4% (17 of 18)	0.01
		Q2: Highest outflow?	95% (19 of 20)	83.3% (15 of 18)	1.37
		Q3: Most in stock?	75% (15 of 20)	38.9% (7 of 18)	**5.07***
		Q4: Least in stock?	80% (16 of 20)	33.3% (6 of 18)	**8.46****
	Not similar	Q1: Highest inflow?	100% (12 of 12)	76.9% (10 of 13)	3.15
		Q2: Highest outflow?	100% (12 of 12)	84.6% (11 of 13)	2.01
		Q3: Most in stock?	50% (6 of 12)	46.2% (6 of 13)	0.03
		Q4: Least in stock?	58.3% (7 of 12)	46.2% (6 of 13)	0.37

*$p<0.05$; **$p<0.01$.

responses in question 4 was greater (80 percent) when participants compared problems similar in both behavior and surface to the DS problem than when they compared problems not similar in both (33.3 percent), $\chi^2(1)=8.46$, $p<0.01$. Thus Hypothesis 2 is also supported: the effect of behavioral similarity depends on surface similarity, and the advantage of similarity occurs only when students compared problems that were similar in both behavioral and surface characteristics to the DS task.

Summary

The finding that comparing problems behaviorally similar to the DS task resulted in more accurate responses in the subsequent DS problem is an important one. The process of comparison, as suggested by the analogy literature, is helpful in finding the common degree of overlap in the relationship between objects (Gentner, 1983). A possible concern is that the step-by-step instructions given to participants to interpret the graphs and to draw the stock function may have contributed to the improved accuracy of responses in the DS problem, rather than the comparison of the analogical problems. However, the fact that the results show improved responses in the DS problem, particularly for those participants that incorrectly predicted the shape of the stock during the instructions, suggest that it is the comparison process and not the given instructions that are leading to improved accuracy on the DS problem. Having students compare behaviorally similar problems helped them to answer question 4 in the subsequent DS task more accurately.

Our analyses indicate that the effect of comparing two behaviorally similar problems depends on their surface similarity as well. That is, our results cannot help us draw definite conclusions on the independent effects of behavioral and surface similarity. Our findings from the Experiment 1 suggest that for students to see the behavioral relations between two SF problems they also need to have similar superficial attributes, even though these are irrelevant to understanding the stock over time. It is important to determine whether participants are able to resolve the DS task on the basis of comparing corresponding behavioral similarity only, or if the effect is due to mere irrelevant appearance in the graphs. The ability to see the behavioral similarity of two problems is essential to adapting to novel contexts (Holyoak and Thagard, 1995). Thus it is necessary to disentangle the effects of behavioral similarity from those of surface similarity. In addition, some of the experimental groups have less participants than other groups. The low *n* in this experiment may have an impact on the explanatory power of these results. Although we only report statistically significant results, the effect of size is not easily interpretable for the non-parametric statistical tests needed for these analyses. Thus, in general, increasing the sample size would help accomplish greater confidence in our results. These issues are addressed in Experiment 2.

Experiment 2: Separating the effects of surface and behavioral similarity

Empirical studies have shown that greater surface similarities may enhance the effect of behavioral similarity (Gick and Holyoak, 1983). Furthermore, according to structure-mapping theory (Gentner, 1983), the process of comparison first relies on surface object matches and then combines with the behavioral alignment. Thus an important question to determine is the effect of behavioral similarity independent of surface similarity. In our

DOI: 10.1002/sdr

Experiment 1, we used two identical problems during the comparison process. The comparison of two identical problems may have highlighted the surface commonalities rather than the behavioral commonalities between the problems.

Thompson *et al*. (2000) suggest that focusing on problems with different surface features may promote the abstraction of the common behavioral features of problems. In a realistic experiment of negotiation, they employed two comparison cases with little surface similarity. Making a comparison between cases that lacked surface similarity resulted in less competition with the recognition of relevant behavioral similarity, increased the encoding of relevant behavioral information, and ended up in more successful subsequent negotiation.

Here we test the effect that the presence and lack of surface similarity in the comparison process would have on the abstraction of behavioral features, and thus on success in the subsequent DS task. We presented participants with two comparison problems that varied in relation to *each other as well as to the DS problem* in surface and behavioral similarity. Again, our goal was to test the beneficial effect of behavioral similarity (Hypothesis 1), but also to help disentangle the effect of surface similarity. Thus, we hypothesized that:

$H_{3.}$ Participants that compare two different problems with behavioral similarity to each other and to the DS problem would respond more accurately to stock questions in the DS problem, regardless of surface similarity.

Experimental design

This experiment again involved a comparison phase followed by the DS task. For the comparison phase, participants were randomly assigned to one of four experimental conditions (Figure 3). Participants were asked to compare two problems that varied in behavioral similarity (Similar and Not Similar) and surface (Similar and Not Similar) to each other and to the DS problem: problems were similar in behavioral characteristics and similar in surface (B & S), similar in behavioral characteristics and not similar in surface (B & NOT S), not similar in behavioral characteristics and similar in surface (NOT B & S), or not similar in behavioral characteristics and not similar in surface (NOT B & NOT S). For an example of the full instructions of B & NOT S in Experiment 2, see Appendix B.

In the B & S condition, both graphs in each of the two comparison problems are identical to each other and identical to the graph seen in the DS task, making all three graphs both similar in behavioral characteristics and surface. In B & NOT S, both comparison graphs are similar in behavioral characteristics to each other and to the DS task because the resulting stock curve is the same inverted U-shaped curve as in the DS task. The problems are not similar in surface because both comparison graphs display different time periods from each other and the DS graph (19 time periods vs. 14 time periods vs. 30 time periods). In NOT B & S, both comparison problems are not similar in behavioral characteristics with each other and to the DS task because each graph produces a different stock curve (linearly decreasing curve vs. linearly increasing curve vs. inverted U-shaped curve). All the graphs in this condition are similar in surface because they all display the inflow and outflow over 30 time periods. In NOT B & NOT S, both graphs are not similar in behavioral characteristics with each other and to the DS task because each graph produces a different stock curve (an increasing curve that evens out vs. a logarithmically decreasing curve vs. the inverted U-shaped curve), and all the graphs are not similar in surface because they each display the inflow and outflow over varying time periods (15 time periods vs. 10 time periods vs. 30 time periods).

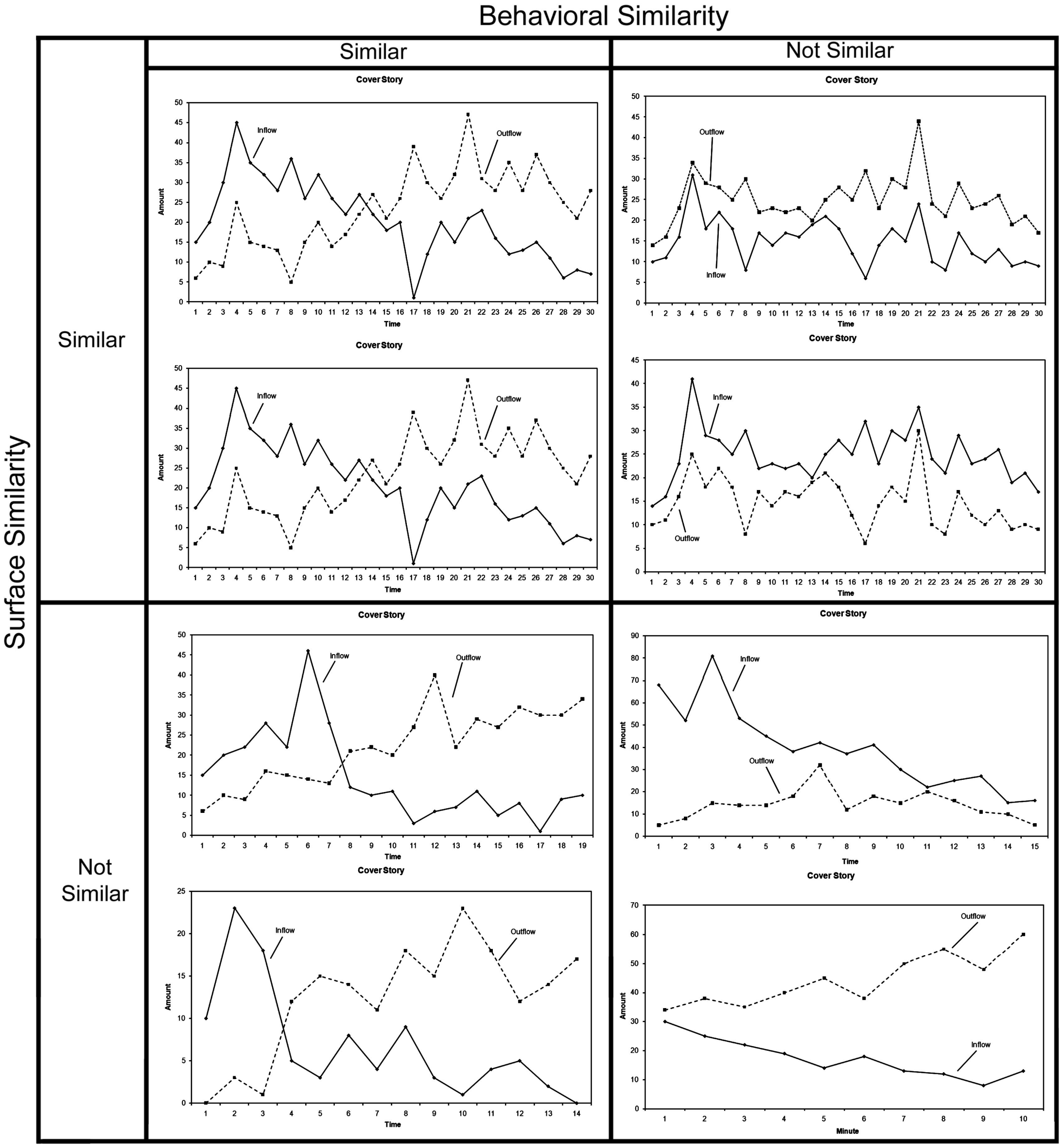


Fig. 3. Experimental design for Experiment 2. Participants were given two graphs corresponding to one of the four conditions: similar in behavioral characteristics and similar in surface (B & S); similar in behavioral characteristics and not similar in surface (B & NOT S); not similar in behavioral characteristics and similar in surface (NOT B & S); and not similar in behavioral characteristics and not similar in surface (NOT B & NOT S). The two graphs were given under two different cover stories (the bank account and the water tank)

The main dependent variable was the proportion of correct responses to the four questions in the subsequent DS problem.

DOI: 10.1002/sdr

Participants

Participants were recruited from local universities using electronic boards and a website (N=83). The average age of this group was 23.57 years (SD=5.54), ranging from 18 to 56. Over half of the participants (61.4 percent) were male and 51 percent of participants majored in mechanical and electrical engineering and other sciences. As in Experiment 1, participants were randomly assigned to one of the four conditions (B & S: N=19; B & NOT S: N=23; NOT B & S: N=21; NOT B & NOT S: N=20).

Procedure

During the comparison phase, individuals were given two problems that were either different or similar in surface and behavioral characteristics to each other according to the assigned condition, each with instructions and a graph showing the *Flows* of the problem (Figure 3). Again, one of the problems used the *water* cover story and the other used the *bank* cover story. Instructions and procedures were identical to those of Experiment 1. After the comparison phase, participants were given the DS problem and asked to answer the four questions.

Results

Comparison phase

Similar to the results of Experiment 1, we found 69.9 percent of participants drew the stock correctly in at least one comparison problem. The proportion did not vary considerably by condition: B & S (68 percent); B & NOT S (74 percent); NOT B & S (66 percent); and NOT B & NOT S (70 percent); again indicating they were of similar difficulty (Cronin *et al.*, 2009). Of those that drew the stock incorrectly in either or both comparison problems, 56 percent drew a shape identical to the inflow or net flow, thus following the correlation heuristic. The proportions of incorrect responses that followed the correlation heuristic per condition are: 33.3 percent (B & S); 66.7 percent (B & NOT S); 71.4 percent (NOT B & S); and 50 percent (NOT B & NOT S). Thus fewer participants followed the correlation heuristic in all but the B & NOT S condition.

The DS problem

Table 4 presents the raw percent of correct responses to the DS problem for each question in each of the four conditions. In general, 95.2 percent, 95.2 percent, 62.7 percent, and 60.2 percent of participants answered questions 1, 2, 3, and 4 correctly, respectively. These proportions are comparable to previous studies, but this experiment shows slightly higher proportions of correct responses compared to those of Experiment 1 in general. The predominant types of error in questions 3 and 4 are similar to those in Experiment 1: 18 out of 31 errors made in question 3 were the maximum net inflow and 11 out of 33 errors made in question 4 were the maximum net outflow. Again, the next most frequent type of error made was the "cannot be determined" response to questions 3 and 4.

The independent effects of surface and behavioral similarities are shown in Table 5. These results indicate that overall only surface similarity makes a difference in the proportion of accurate responses to question 4, $\chi^2(1)=7.02$, $p<0.01$. The proportion of correct responses was higher (75 percent) in the DS problem when there was surface similarity than when the surface of the problems was not similar (46.5 percent).

DOI: 10.1002/sdr

Table 4. Distributions of responses to the four questions in the DS problem after the comparison in each of the four conditions in Experiment 2. The top part of the table shows the proportions of correct responses in each question and condition; the shaded cells contain the frequencies of correct answers to each of the four questions. The bottom part of the table shows the proportion of errors of different types

	Q1: Most entering?				Q2: Most leaving?				Q3: Most in store?				Q4: Fewest in store?			
Type of response	B & S	B & NOT S	NOT B & S	NOT B & NOT S	B & S	B & NOT S	NOT B & S	NOT B & NOT S	B & S	B & NOT S	NOT B & S	NOT B & NOT S	B & S	B & NOT S	NOT B & S	NOT B & NOT S
Correct responses																
Q1: Most entering: t=4	18 94.7%	22 95.7%	20 95.2%	19 95%	0	0	0	0	0	1 4.3%	0	1 5%	0	0	0	1 5%
Q2: Most leaving: t=21	0	0	0	0	18 94.7%	22 95.7%	20 95.2%	19 95%	0	0	0	0	0	0	2 9.4%	1 5%
Q3: Most in store: t=13	0	0	0	0	0	0	0	0	14 73.7%	13 56.6%	13 61.9%	12 60%	0	0	1 4.8%	0
Q4: Fewest in store: t=30	0	0	0	0	0	0	0	0	0	1 4.3%	0	1 5%	15 78.9%	12 52.2%	15 71.4%	8 40%
Errors																
Max. net inflow (entering − leaving): t=8	1 5.3%	1 2.15%	1 4.8%	1 5%	0	0	0	0	3 15.8%	7 30.5%	5 23.8%	3 15%	0	2 8.7%	0	0
Max. net outflow (leaving − entering): t=17	0	0	0	0	1 5.3%	1 4.3%	1 4.8%	1 5%	0	0	0	1 5%	1 5.3%	5 21.7%	1 4.8%	4 20%
Initial in store t =1	0	0	0	0	0	0	0	0	0	0	2 9.5%	0	0	2 8.7%	0	3 15%
Can't be determined	0	0	0	0	0	0	0	0	2 10.5%	1 4.3%	1 4.8%	2 10%	3 15.8%	2 8.7%	1 4.8%	3 15%
Other	0	1 2.15%	0	0	0	0	0	0	0	0	0	0	0	0	1 4.8%	0
Total	19	23	21	20	19	23	21	20	19	23	21	20	19	23	21	20

DOI: 10.1002/sdr

Table 5. Experiment 2. Analyses of the effects of surface and behavioral similarities on the percentage of correct responses to the DS problem. Statistical differences between correct and incorrect responses are in bold

		Surface			Behavioral		
		N=63			N=63		
	Questions	Similar	Not similar	$\chi^2(1)$	Similar	Not similar	$\chi^2(1)$
Overall	Q1: Highest inflow?	95% (38 of 40)	95.3% (41 of 43)	0.01	95.2% (40 of 42)	95.1% (39 of 41)	0.001
	Q2: Highest outflow?	95% (38 of 40)	95.3% (41 of 43)	0.01	95.2% (40 of 42)	95.1% (39 of 41)	0.001
	Q3: Most in stock?	67.5% (27 of 40)	58.1% (25 of 43)	0.78	64.3% (27 of 42)	61% (25 of 41)	0.10
	Q4: Least in stock?	75% (30 of 40)	46.5% (20 of 43)	**7.02****	64.3% (27 of 42)	56.1% (23 of 41)	0.59
Correct stock shape	Q1: Highest inflow?	96.3% (26 of 27)	93.5% (29 of 31)	0.22	93.3% (28 of 30)	96.4% (27 of 28)	0.28
	Q2: Highest outflow?	96.3% (26 of 27)	93.5% (29 of 31)	0.22	93.3% (28 of 30)	96.4% (27 of 28)	0.28
	Q3: Most in stock?	74.1% (20 of 27)	67.7% (21 of 31)	0.28	63.3% (19 of 30)	78.6% (22 of 28)	1.62
	Q4: Least in stock?	77.8% (21 of 27)	51.6% (16 of 31)	**4.28***	60% (18 of 30)	67.9% (19 of 28)	0.39
Incorrect stock shape	Q1: Highest inflow?	92.3% (12 of 13)	100% (12 of 12)	0.96	100% (12 of 12)	92.3% (12 of 13)	0.96
	Q2: Highest outflow?	92.3% (12 of 13)	100% (12 of 12)	0.96	100% (12 of 12)	92.3% (12 of 13)	0.96
	Q3: Most in stock?	53.8% (7 of 13)	33.3% (4 of 12)	1.10	66.7% (8 of 12)	23.1% (3 of 13)	**4.81***
	Q4: Least in stock?	69.2% (9 of 13)	33.3% (4 of 12)	3.22	75% (9 of 12)	30.8% (4 of 13)	**4.90***

*$p<0.05$; **$p<0.01$.

When breaking down the results according to the accuracy with which the stock was drawn during the comparison phase, we found that surface similarity was significant for those that drew the stock shape correctly, $\chi^2(1)=4.28$, $p<0.05$. The proportion of correct responses to question 4 was greater (77.8 percent) when participants compared problems with surface similarity than when they compared problems without surface similarity (51.6 percent). Thus, after having drawn the stock correctly during comparison, participants were able to respond more accurately to question 4 in the DS problem having compared problems similar in surface than when they compared problems not similar in surface.

In contrast, behavioral similarity was significant for those that drew the stock shape incorrectly; they were able to respond more accurately to questions 3 and 4 in the DS problem after having compared problems that shared behavioral similarity than after having compared behaviorally not similar problems. The proportion of correct responses to question 3 was greater (66.7 percent) when participants compared problems with behavioral similarity than when they compared problems not similar in behavior (23.1 percent), $\chi^2(1)=4.81$, $p<0.05$. Similarly, the proportion of correct responses to question 4 was greater (75 percent) when participants compared problems that were similar in behavior than when they compared problems not similar in behavior (30.8 percent), $\chi^2(1)=4.90$, $p<0.05$. Analyses of surface and behavioral similarity together showed no significant interaction effects.

Summary

These results provide support for Hypothesis 3: the comparison of problems that are similar in surface helps those that drew the stock correctly during comparison, while the comparison of problems that are similar in deep behavior helps those that incorrectly predicted the stock's shape. These results support and extend those of Experiment 1. Now it was possible to separate the effects of surface and behavioral similarity, indicating when each type of similarity can be beneficial to judging the stock in the subsequent DS problem. This result supports the findings of Thompson *et al*. (2000): that comparing problems that are different in surface helps identify the common behavioral features of the problems and improves the responses in the subsequent problem. However, our results also clarify this finding further; it is clear that comparing problems that are similar in surface also helps enhance the behavioral similarities for those that are able to identify these similarities and draw the stock correctly during comparison. Thus surface similarity can be beneficial too once the behavioral similarities are understood.

Discussion

The SF failure is a robust problem that may need concrete interventions to highlight the relationship between flows and their effect on a stock over time. Many decisions in the real world are based on the understanding of accumulation, and we would argue that many real-world individual, organizational, and social problems are rooted in the misunderstanding of these basic processes of dynamic systems. A main problem is that we understand little of what are the mental processes that lead to misunderstandings and errors in the judgments of stock levels. Important progress has been made in identifying the "correlation heuristic" (Cronin *et al*., 2009) as a way in which people reason in these problems, but explanations as to why people often assume that the stock behaves like the flows and how to help such erroneous reasoning are needed.

This research contributes to our understanding of how analogical comparisons influence the judgments of SF problems and the use of the correlation heuristic. Results from Experiment 1 indicated that comparing problems with similar behavioral features facilitates better performance in the DS problem; but the results also indicated that the benefit of behaviorally similar comparisons was dependent on the surface similarity as well. This result highlights our human limitations to see the common behavioral characteristics of problems that differ in their surface characteristics. This cognitive limitation supports existing literature that suggests that the problems used during the comparison process need to share surface and behavioral characteristics with the test problem in order to be effective (Holyoak and Koh, 1987). This result seemed disappointing, as it essentially suggests a practice effect: in the B & S condition, people made judgments and analyzed problems that were essentially identical (except for the cover story). After examining some of the written comparisons from participants in the B & S condition, it is clear that most participants saw the two comparison problems as "essentially the same", and thus the positive transfer to the DS problem may be due to a practice effect. The effect of practice and experience has been found to influence the accuracy of responses in the DS problem (Cronin *et al*., 2009; Brunstein *et al*., 2010), and this effect is interesting in its own right. However, we expected that through the process of analogical comparison participants would be able to see the similarity of the behavioral features of problems, beyond surface similarity. This kind of reasoning ability is what will help people in the real world to solve SF problems across contexts

DOI: 10.1002/sdr

and domains that are analogical in the behavioral structure, even when they are different in the surface features.

Fortunately, our findings from Experiment 1 also indicate that comparing two behaviorally similar problems to the DS task was particularly beneficial to those participants that *incorrectly* predicted the shape of the stock during comparison. Thus we believe that the analogical comparison process of two behaviorally similar problems to the DS problem helped those that did not understand the relationships to the target problem in the first place. This is an important observation because it highlights the benefits of analogical reasoning in the instruction and education of system dynamics concepts. Analogical reasoning has been considered key to learning abstract concepts and procedures, and to improving our ability to transfer representations across contexts (Novick, 1988; Holyoak and Thagard, 1995; Richland *et al.*, 2004). In reasoning with SF problems, it is important to understand the underlying behavioral characteristics of the problems, the connections and relations within the parts of the problems, and the goals of the objects—even when the superficial features of the problems are not necessarily similar. Our results indicate that the process of analogical comparison can achieve this goal, particularly helping those that do not understand the behavioral relationships of SF problems in the first place.

The results from Experiment 2 further address the benefits of comparing problems that share behavioral similarity, independently from surface similarity. Our results indicate that surface similarity is important for those that correctly drew the stock in the comparison task: when individuals already understood the behavioral similarity of the problems during comparison, the surface similarity helped reduce the SF failure and the use of the correlation heuristic in the subsequent DS task. Most importantly, our results suggest a way to best promote behavioral understanding of SF problems: the comparison of two different problems that are similar to the DS task reduced the use of the correlation heuristic in the subsequent DS task. Thus important underlying characteristics of SF problems can be learned by comparing different problems that are similar in their underlying behavioral relationships to the target problem.

As we suggest above, our results have direct implications to the instruction and education of system dynamics concepts. Formal system dynamics courses often employ causal-loop diagrams and stock and flow diagrams to introduce the concepts of accumulation and flows. These types of diagrams are used in introductory courses and in the communication of system dynamics concepts. It has often been assumed that these diagrams are simple and easily communicate the concepts (Richardson, 1986), but these instruments have many limitations (Lane, 2008). For example, causal-loop diagrams are imprecise at pointing out the distinctions between stocks and flows (Richardson, 1986; Lane, 2008), and stock and flow diagrams are obscure at representing and explaining a system's behavior (Lane, 2008).

Pala and Vennix (2004) performed tests using the DS task before and after participants took an introductory system dynamics course. Although they report a significant effect of the course instruction on the improvement of answers to questions 3 and 4 of the DS task, the proportion of correct responses in the after test remained disappointing: 60 percent and 45 percent for questions 3 and 4 respectively (Pala and Vennix, 2004). Similarly, a recent study shows the effectiveness of formal training using a pre-test and post-test after instruction with the DS task (Sterman, 2010). An experiment with graduate students taking an introductory course in system dynamics resulted in large improvements in people's understanding of accumulation and reduction in the use of the correlation heuristic. Although this represents an important step towards understanding the benefit of formal instruction, more formal assessments are needed (Pala and Vennix, 2004). We believe that more control over what concepts and instructional strategies produce a positive effect on people's understanding is required. The results of the current experiment suggest one way in which instruction and training of system dynamics concepts could be improved. Our results demonstrate that by the process of comparing different problems that are behaviorally similar to the DS problem, those that drew the stock incorrectly during the comparison process obtained 66.7 percent and 75 percent of correct responses in questions 3 and 4 respectively, compared to those that compared problems that were not behaviorally similar to the DS problem (23.1 percent and 30.8 percent of correct responses respectively). Thus analogical comparison may be one instructional strategy for remedying the SF failure in adults. The use of analogical comparison through different problems that are similar in the behavioral characteristics to the target problem will help those that do not understand SF concepts in the first place. Also, the use of analogical comparisons of will help those that already understand SF concepts.

An interesting limitation and a good potential opportunity for future research is the fact that analogical reasoning is highly dependent on the level of experience in a particular domain. For example, Chi *et al*. (1981) found that those with less experience in the task domain based their judgments mostly on surface features, while those with more experience in a domain relied more on behavioral similarity. They demonstrated this effect in the physics domain. However, Brunstein *et al*.'s results (2010) suggest that domain knowledge might not be enough to see the underlying similarity of SF problems. Medical students performed equally poorly as undergraduates with no medical knowledge in problems that required medical domain experience. Thus a follow-up study may involve populations with different types of experience—for example, students experienced in system dynamics concepts as compared to those that are less experienced. Furthermore, future research should engage in formal analysis to learn what basic components of the SF problem people acquire through comparison. It would also be important to test how well people retain the information learned from comparisons after some time has passed. Would improvements gained by the comparison process be long lasting? And would participants that are given a similar task a week later perform equally as well? These and similar questions are the topic of future research.

Notes

1. The Algebra Standard of the U.S.A. indicates that concepts of accumulation and rate of change should be introduced in grades 6–8. Students should be able to examine the relationships between stocks and flows depicted in graphs and understand the differences and relationships between them.
2. The concept of behavioral similarity is commonly referred to as "structure similarity" in the psychology literature. However, given that in system dynamics "structure" refers to the interaction of the feedback loops and this may cause confusion in the use of the term, we decided to refer to the term "structure similarity" as "behavioral similarity" throughout this paper.

Acknowledgements

This research was partly supported by the National Science Foundation (Human and Social Dynamics: Decision, Risk, and Uncertainty, Award No. 0624228) award to Cleotilde Gonzalez. We thank Polina Vanyukov for help in running the experiments reported in this paper.

Biographies

Cleotilde Gonzalez is an Associate Research Professor in the Department of Social and Decision Sciences at Carnegie Mellon University. Her research focuses on cognitive aspects of decision making in dynamic environments. She uses behavioral, computational, and functional brain-imaging approaches to understand how people make decisions in dynamic, complex environments. She is the founder and director of the Dynamic Decision Making Laboratory at Carnegie Mellon (http://www.cmu.edu/ddmlab), which currently employs several postdoctoral fellows, researchers, and programmers.

Hau-yu Wong is a Research Associate with the Dynamic Decision Making Laboratory at Carnegie Mellon University. She previously earned a Bachelor of Science in psychology and international relations from Carnegie Mellon.

References

Ben-Zeev T, Star JR. 2001. Spurious correlations in mathematical thinking. *Cognition and Instruction* **19**(3): 253–275.

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

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

Chi MTH, Feltovich PJ, Glaser R. 1981. Categorization and representation of physics problems by experts and novices. *Cognitive Science* **5**(2): 121–152.

Cronin M, Gonzalez C. 2007. Understanding the building blocks of system dynamics. *System Dynamics Review* **23**(1): 1–17.

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

Gentner D. 1983. Structure-mapping: a theoretical framework for analogy. *Cognitive Science* **7**: 155–170.

Gentner D, Markman AB. 1997. Structure mapping in analogy and similarity. *American Psychologist* **52**: 45–56.

Gick ML, Holyoak KJ. 1983. Schema induction and analogical transfer. *Cognitive Psychology* **15**(1): 1–38.

Harel G, Behr M, Post T, Lesh R. 1992. The blocks task: comparative analyses of the task with other proportion tasks and qualitative reasoning skills of 7th-grade children in solving the task. *Cognition and Instruction* **9**(1): 45–96.

Holyoak KJ, Koh K. 1987. Surface and structural similarity in analogical transfer. *Memory and Cognition* **15**(4): 332–340.

Holyoak KJ, Thagard P. 1995. *Mental Leaps: Analogy in Creative Thought*. MIT Press: Cambridge, MA.

DOI: 10.1002/sdr

Kurtz KJ, Miao C, Gentner D. 2001. Learning by analogical bootstrapping. *Journal of Learning Sciences* **10**(4): 417–446.

Lane DC. 2008. The emergence and use of diagramming in system dynamics: a critical account. *Systems Research and Behavioral Science* **25**(1): 3–23.

Lowenstein G, Thompson LL, Gentner D. 1999. Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin and Review* **6**(4): 586–597.

Medin DL, Goldstone RL, Markman AB. 1995. Comparison and choice: relations between similarity processing and decision processing. *Psychonomic Bulletin and Review* **2**(1): 1–19.

Novick LR. 1988. Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology. Learning, Memory, and Cognition* **14**(3): 510–520.

Pala O, Vennix JAM. 2004. Effects of system dynamics education on system thinking inventory task performance: *System Dynamics Review* **21**(2): 139–162.

Richardson GP. 1986. Problems with causal-loop diagrams. *System Dynamics Review* **2**: 158–170.

Richland LE, Holyoak KJ, Stigler JW. 2004. Analogy use in eighth-grade mathematics classrooms. *Cognition and Instructions* **22**(1): 37–60.

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

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

Thompson WL, Gentner D, Loewenstein J. 2000. Avoiding missing opportunities in managerial life: analogical training more powerful than individual case training. *Organizational Behavior and Human Decision Processes* **82**(1): 60–75.

Van Dooren W, De Bock D, Depaepe F, Janssens D, Verschaffel L. 2003. The illusion of linearity: expanding the evidence towards probabilistic reasoning. *Educational Studies in Mathematics* **53**: 113–138.

Van Dooren W, De Bock D, Hessels A, Janssens D, Verschaffel L. 2004. Remedying secondary school students' illusion of linearity: a teaching experiment aiming at conceptual change. *Learning and Instruction* **14**: 485–501.

Van Dooren W, De Bock D, Hessels A, Janssens D, Verschaffel L. 2005. Not everything is proportional: effects of age and problem type on propensities for overgeneralization. *Cognition and Instruction* **23**(1): 57–86.x

Appendix A: Example of full instructions given in the comparison task for Experiment 1

Problem 1 of the comparison task for condition B & NOT S

A factory storage tank is being filled and emptied by a number of different pumps. The left graph below shows the amount of water flowing into the tank (Inflow) and flowing out of the tank (Outflow) per minute over a period of 30minutes.

Please do the activities described in the next paragraph using the Water Flow (left) graph:

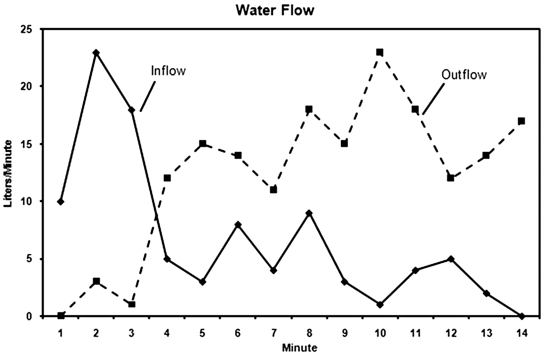
1 The maximum rate of water flow into the tank occurred in the 2nd minute. The maximum rate of water flow out of the tank occurred in the 10th minute. **Please CIRCLE those points**.

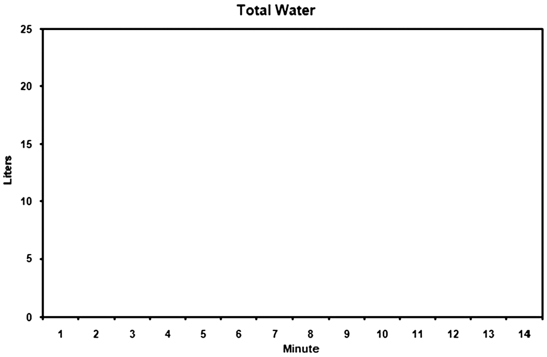
2 During the period from the 1st minute to the 3rd minute the rate of Inflow was greater than the rate of Outflow. The total amount of water in the tank increased during this time period. During the period from the 3rd minute to the 14th minute the rate of

Outflow was greater than the rate of Inflow. The total amount of water in the tank decreased during this time period. **Please mark the difference between the Inflow and the Outflow in these two sections of the graph**.

3 The maximum amount of water in the tank was in the 3rd minute. The minimum amount of water in the tank was in the 14th minute. **Please CIRCLE these points**.

In the space provided on the right, please draw the behavior of the tank's total amount of water during this time period. Do not worry about the exact quantities. We are interested in your drawing of the shape of the curve.





Problem 2 of the comparison task for condition B & NOT S—this was provided on a different page

Money is being deposited and withdrawn from a bank account. The graph on the left shows the amount of money deposited into and withdrawn from the account per day over a period of 30 days.

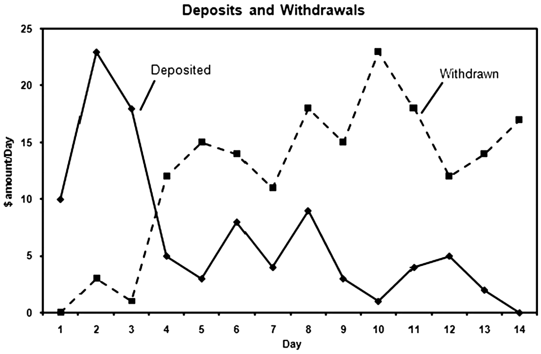
Please do the activities described in the next paragraph using the Deposits and Withdrawals (left) graph:

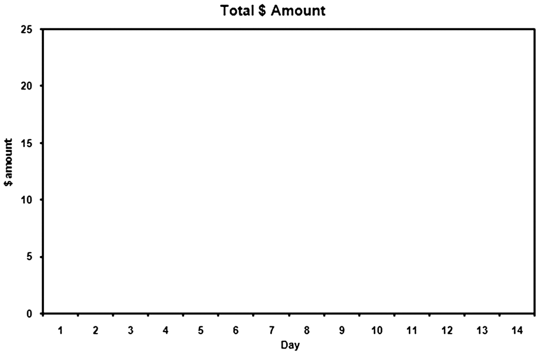
1 The maximum amount of money per day deposited into the account occurred on the 2nd day. The maximum amount of money per day withdrawn from the account occurred on the 10th day. **Please CIRCLE those points**.

2 During the period from the 1st day to the 3rd day the amount of money deposited into the account per day was greater than the amount of money withdrawn from the account per day. The total amount of money in the account increased during this time period. During the period from the 3rd day to the 14th day the amount of money withdrawn from the account per day was greater than the amount of money deposited in the account per day. The total amount of money in the account decreased during this time period. **Please mark the difference between the Inflow and the Outflow in these two sections of the graph**.

3 The maximum amount of money in the account was on the 3rd day. The minimum amount of money in the account was on the 14th day. **Please CIRCLE these points**.

In the space provided on the right, please draw the behavior of the total amount of money in the account during this time period. Do not worry about the exact quantities. We are interested in your drawing of the shape of the curve.





In the space below describe how the two problems are similar. Please use concrete details in describing the similarity.

Appendix B: Example of full instructions given in the comparison task for Experiment 2

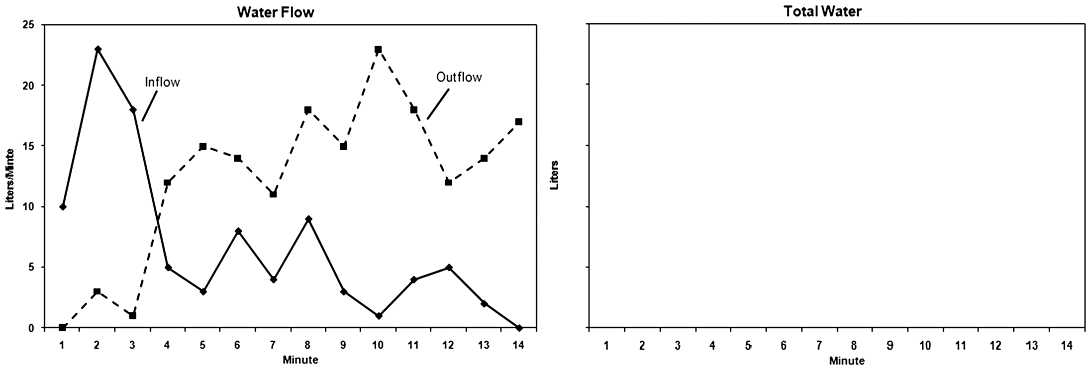
Problem 1 of the comparison task for condition B & NOT S

A factory storage tank is being filled and emptied by a number of different pumps. The left graph below shows the amount of water flowing into the tank (Inflow) and flowing out of the tank (Outflow) per minute over a period of 14 minutes.

Please do the activities described in the next paragraph using the Water Flow (left) graph:

1 The maximum rate of water flow into the tank occurred in the 2nd minute. The maximum rate of water flow out of the tank occurred in the 10th minute. **Please CIRCLE these points**.
2 During the period from the 1st minute to the 3rd minute the rate of Inflow was greater than the rate of Outflow. The total amount of water in the tank increased during this time period. During the period from the 4th minute to the 14th minute the rate of Outflow was greater than the rate of Inflow. The total amount of water in the tank decreased during this time period. **Please mark the difference between the Inflow and the Outflow in these two sections of the graph**.
3 The maximum amount of water in the tank was on the 3rd day. The minimum amount of water in the tank was on the 14th day. **Please CIRCLE these points**.

In the space provided on the right, please draw the behavior of the tank's amount of water during this time period. Do not worry about the exact quantities. We are interested in your drawing of the shape of the curve.



Problem 2 of the comparison task for condition B & NOT S—this was provided on a different page

Money is being deposited and withdrawn from a bank account. The graph on the left shows the amount of money deposited into and withdrawn from the account per day over a period of 14 days.

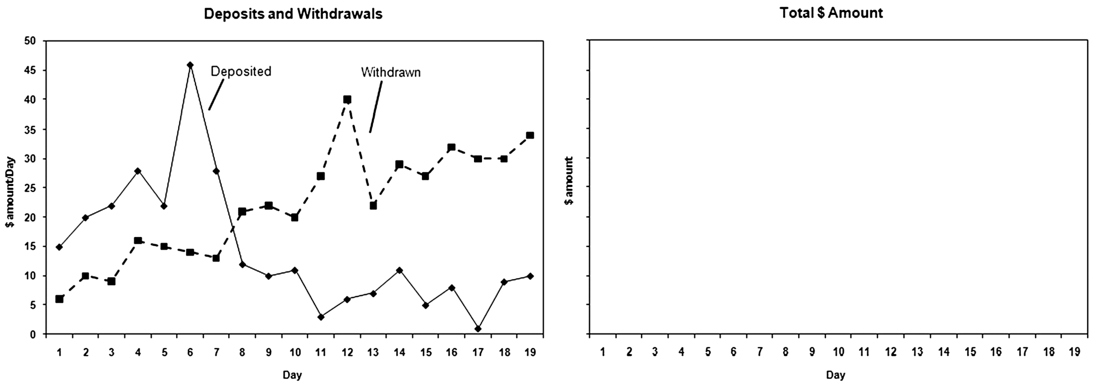
Please do the activities described in the next paragraph using the Deposits and Withdrawals (left) graph:

1 The maximum amount of money per day deposited into the account occurred on the 6th day. The maximum amount of money per day withdrawn from the account occurred on the 12th day. **Please CIRCLE these points**.

2 During the period from the 1st day to the 7th day the amount of money deposited into the account per day was greater than the amount of money withdrawn from the account per day. The total amount of money in the account increased during this time period. During the period from the 8th day to the 19th day the amount of money withdrawn from the account per day was greater than the amount of money deposited into the account per day. The total amount of money in the account decreased during this time period. **Please mark the difference between the Deposited amounts per day and the Withdrawn amounts per day in these two sections of the graph**.

3 The maximum amount of money in the account was on the 7th day. The minimum amount of money in the account was on the 19th day. **Please CIRCLE these points**.

In the space provided on the right, please draw the behavior of the total amount of money in the account during this time period. Do not worry about the exact quantities. We are interested in your drawing of the shape of the curve.



In the space below describe how the two problems are similar. Please use concrete details in describing the similarity.