The Dual-Process Theory and Understanding of Stocks and Flows

Arash Baghaei Lakeh (corresponding author)
PhD Candidate
Industrial & Systems Engineering
Virginia Tech
Blacksburg, VA, USA
arashb@vt.edu

Navid Ghaffarzadegan, PhD
Assistant Professor
Industrial and Systems Engineering
Virginia Tech
Blacksburg, VA, USA
navidg@vt.edu

Abstract:

Recent evidence suggests that using the System-2 mode of thinking can improve people's performance in stock-flow (SF) tasks. In this paper, we further investigate the effects by implementing several different interventions in two studies. First, we replicate a previous finding that answering analytical questions before the SF task about doubles the likelihood of answering the stock questions correctly. We also investigate effects of three other interventions that can potentially prime participants to use their System-2. Specifically, the first group is asked to justify their response to the SF task; the second group is warned about the difficulty of the SF task; and the third group is offered information about cognitive biases and the role of the analytic mode of thinking. We find that the second group showed a statistically significant improvement in their performance. We claim that there are simple interventions which can modestly improve people's response in SF tasks.

Keywords:

Accumulation, Stock-Flow Failure, Dual Process Theory, Cognitive Capacity, Analytical Thinking, System 2 Thinking, Experimental design

INTRODUCTION

Several studies have shown that people have difficulty inferring the behavior of a stock variable based on information about its inflows and outflows (Booth Sweeney and Sterman, 2000; Kapmeier, 2004; Cronin *et al.*, 2009; Sterman, 2010; Abdel-Hamid *et al.*, 2014; Qi and Gonzalez, 2015). However, little progress has been made on how to improve people's understanding of stocks and flows (SF) and many interventions that have been designed for such a purpose have not been successful. Altering the representation of the task, providing different contexts for an SF task, and increasing participants' motivation have all failed to significantly improve people's performance in SF tasks (Cronin *et al.*, 2009).

Education is shown to be one of the few, robust, successful interventions for improving people's performance in SF tasks (Kainz and Ossimitz, 2002; Pala and Vennix, 2005; Sterman 2010; Gonzalez and Wong, 2009). The core idea is to teach people about stocks and flows. For example, Sterman (2010) shows that an introductory class of system dynamics will improve the rate of correct answers to an SF task. However, the main limitation of such educational interventions is that teaching is not always feasible, as it is often time consuming and highly resource dependent. The current studies with educational intervention are based on active participation of students during a short or semester-long course (Kains and Ossimitz, 2002; Sterman, 2010). Moreover, the quality of education is teacher dependent (e.g., Darling-Hammond, 2000). Overall, it seems that more studies are needed to further explore other tools and techniques for a *quick* improvement in people's understanding of the concept of accumulation. Our paper moves in this direction, building on a recent idea.

In a recent study, Baghaei Lakeh and Ghaffarzadegan (2015), hereafter referred to as 'BG', develop a low-cost intervention that modestly improves people's performances in an SF task. The authors base their intervention on the idea that the high rate of failure in SF tasks is partially related to how people think about such tasks and that if people use their analytic mode of thinking (also known as the System 2 mode of thinking) (Kahneman and Frederick, 2002), they will be more likely to answer an SF task correctly. They test the hypothesis in an experimental set-up of a control versus a treatment group on Amazon Mechanical Turk (an online crowdsourcing platform commonly used in behavioral studies). BG ask participants in their treatment group an unrelated analytical question prior to the SF task—the hypothesis being that thinking about an analytical question right before the SF task can improve people's performance in the SF task, since it triggers the System 2 mode of thinking. Their results support the hypothesis and demonstrate a modest, yet statistically significant, improvement of responses to SF tasks. An important point in regards to BG's method is that the intervention is a low-cost and quick task, which takes at most a few minutes to implement.

In this article, we build on BG's study and extend their analysis to further investigate the effect of the analytic mode of thinking on SF task performance. We relax the assumption that people

are not familiar with the concept of accumulation (the foundation of educational interventions) and test the hypothesis that more people can provide a correct answer to an SF task if they use their analytic mode of thinking. We present the theoretical foundation of our work in the next section and then report the results of two studies. Our results show that it is possible to modestly improve people's understanding of accumulation by cognitive interventions.

INTUITIVE IMPRESSIONS AND ANALYTIC JUDGMENTS

We base our study on the dual-process theory (Evans, 2003; Kahneman and Frederick, 2002). There are several studies in the fields of reasoning, judgment, and decision-making that suggest that people possess two distinct cognitive systems of thinking known as System 1 and System 2 (for a review refer to Evans, 2003). System 1 (or the 'intuitive mind') refers to the set of cognitive processes that are automatic, effortless, associative, and rapid (parallel), while System 2 (or the 'reflective mind') refers to the cognitive processes that are controlled, effortful, deductive, and slow (serial) (Kahneman and Frederick, 2002).

Kahneman and Frederick (2002) explain how our cognitive systems work together. When facing a cognitive task, System 1 will automatically generate intuitive proposals on how to solve the problem. System 2 can deductively review these proposals and decide whether to reject, approve, or modify them. In some cases, the final answer approved by System 2 might be very close to an intuitive answer offered by System 1. In some other cases, System 2 will dramatically revise the intuitive response by System 1. However, our final judgments are usually highly anchored by the initial impressions generated by System 1.

System 1 operations are crucial in our daily lives, and, even, very complex cognitive procedures can be done through our intuitive mind (Kahneman and Frederick, 2002). However, relying only on System 1 and using heuristics for addressing all cognitive tasks can produce decision-making errors and systematic biases (Tversky and Kahneman, 1974).

The dual-process theory has been applied by scholars in the education studies to investigate why errors occur when people try to solve a problem (e.g., Leron and Hazzan, 2009). There are two general approaches for analyzing errors in problem solving. The first is to assume that errors come from a lack of knowledge; thus, people should be taught to avoid errors (Kainz and Ossimitz, 2002; Pala and Vennix, 2005; and Sterman, 2010). In contrast, the second approach is to analyze errors from a behavioral perspective (BG; Fisher and Gonzalez, 2015; Fischer et al., 2015; Hämäläinen *et al.*, 2013; Gonzalez and Wong, 2012). An example of the latter approach is to consider errors as a kind of cognitive bias caused by the use of heuristics rather than analytic thinking processes. In other words, people judge poorly not necessarily because they are ignorant of the problem, but because they are using cognitive shortcuts (heuristics) to answer the questions. The implication of the latter approach for our problem

(i.e., SF failure) is that it is possible to improve people's performance in accumulation tasks without teaching them about stocks and flows.

The dual-process theory predicts that people will have a better performance in mathematical problems such as SF tasks when they use their System 2 (BG; Leron and Hazzan, 2009). In this paper we sought to empirically evaluate this prediction through experimental settings. To that end, we need to prime people to use their analytic system. Several interventions have been used in past studies to trigger analytic mode of thinking including visual, verbal, contextual, and dysfluency priming methods (Gervais and Norenzayan, 2012; Adam et al., 2007; BG). Generally, people tend to use their System 2 when they face a novel or difficult situation. They are also more likely to use their analytic mode of thinking if they are asked to do so (Louis and Sutton, 1991). Here, we employ BG's intervention along with three other interventions to trigger analytic thinking.

In this article, we propose that SF failure can be partly explained by the dual-process theory. Following BG's study, we hypothesize that people will have a better performance in SF tasks if they are primed to utilize their System 2 mode of thinking. We report two studies. In Study 1, we replicate BG's experiment with a stronger manipulation and a design set-up that is more consistent with Cronin *et al.* (2009). In Study 2, we offer a more in-depth analysis of the effects of using analytic mode of thinking by testing three different interventions of (1) asking participants to justify their answers, (2) warning them about the difficulty of the SF task, and (3) informing them about the two modes of thinking.

STUDY 1

BG's study shows that asking people an unrelated analytical question right before an SF task improves their performance by pushing them to use their System 2 mode of thinking. In study 1, we replicate their results in a similar experimental set up. Here, the hypothesis is that placing SF tasks in the context of analytical questions will expose people to analytic thinking, and this, in turn, improves their performance in SF tasks.

Method

We used Amazon Mechanical Turk as the platform to conduct our experiments. Amazon Mechanical Turk is a web service widely used for crowdsourcing purposes. It has also become an attractive tool for experimental research as it provides researchers easy access to a large pool of subjects and a low cost of performing experiments (Mason and Suri, 2012). The pool of subjects in this platform is significantly more diverse than the typical American college samples that are being used in many psychological experiments. Moreover, it has been shown

that data from Amazon Mechanical Turk are at least as reliable as those obtained via undergraduate or graduate samples (Paolacci et al. 2010; Buhrmester et al. 2011).

We recruited 184 participants for this study. Upon agreeing to participate in our study, participants were randomly distributed in either the control (n=98) or the treatment group (n=86). The self-reported demographics of the participants are reported in Table 1. Each participant was compensated $1 for her or his time after they successfully submitted their work. The compensation was not dependent on the subjects' performance and they only needed to answer all questions in the task to be paid.

Table 1: Demographics of Participants in Study 1

Number of participants	184
Percentage of females	42%
Average age (SD)	31.7 (8.3)
High school education	36.4%
College-level education	51.1%
Graduate-level education	12.5%
Average level of self-reported mathematical expertise* (SD)	2.9 (0.9)

*: On a scale of 1–5; 2 is equivalent to high school mathematics.

In this study, we used the Department Store (DS) task as shown in Figure 1 (Sterman, 2002). Right before the DS task, we ask our participants in the treatment group to answer four questions as reported in Appendix I. These questions were not related to the DS task or the concept of accumulation to avoid any learning effect. The questions were non-trivial, logical, and required a moment of concentration before answering. Before proceeding to the next step, participants in the treatment group needed to answer all of the questions. In line with BG's study, the goal was to set up a cognitive context for the upcoming DS task and make the participants continue using their System 2 mode of thinking.

For the manipulation check, we asked the participants to answer two questions: "Please explain how you came up with the answers to the last two questions in the previous step," and "Explain why you think your approach to answering those questions is correct." We coded their responses in terms of whether they used their System 2 mode of thinking in responding to the DS task. The method of coding was similar to BG's manipulation check.

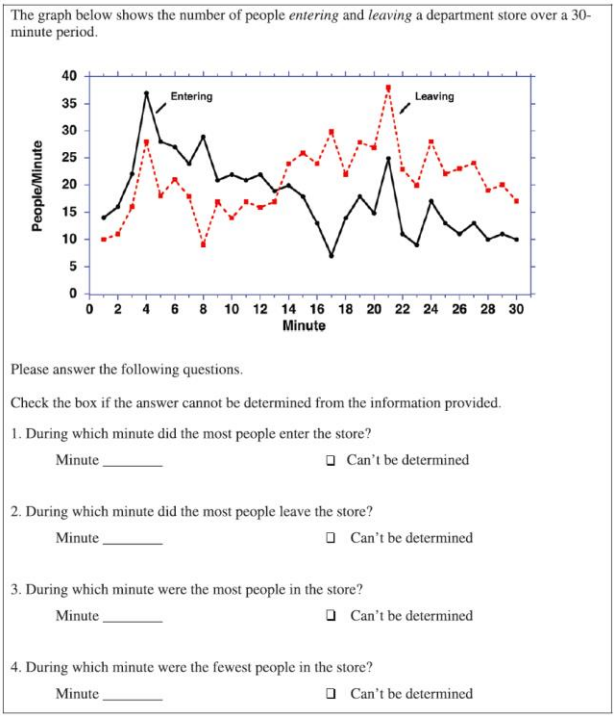


Figure 1. The department store task. (Adopted from Sterman, 2002.)

Study steps were as follows: (1) intervention (only for participants in the treatment group), (2) the DS task, (3) manipulation check questions, and (4) demographic questions (i.e., age, gender, education level, and self-reported mathematical expertise). Steps 2, 3, and 4 were the same for both treatment and control groups.

During the experiment, participants were not able to go back while working on the assignments and only could progress to the next step when they answered all questions of the current step. All steps of the experiment were loaded in one web page. We recorded the number of times that the page was loaded to make sure that it was only loaded once. We did not encounter an instance where the survey was loaded multiple times. Also, participants were informed that they can contact us by email if they have any question. We did not receive any inquiries from the subjects.

In the DS task, the first two questions are about flows and the last two questions are about stocks. The flow questions are asked to make sure that participants could read the graph correctly (Cronin et al., 2009). A strong majority of our participants answered both of these two questions correctly (more than 85% of the participants). In this study, our focus of analysis is on questions 3 and 4 (stock questions), which are coded as shown in Table 2. Specifically, for each participant we defined four variables: Correct Answer, Correlation Heuristic, Basic Error, and Can't be Determined (CBD). These variables count the number of answers to the stock questions that fall in each of the groups shown in Table 2. Correlation

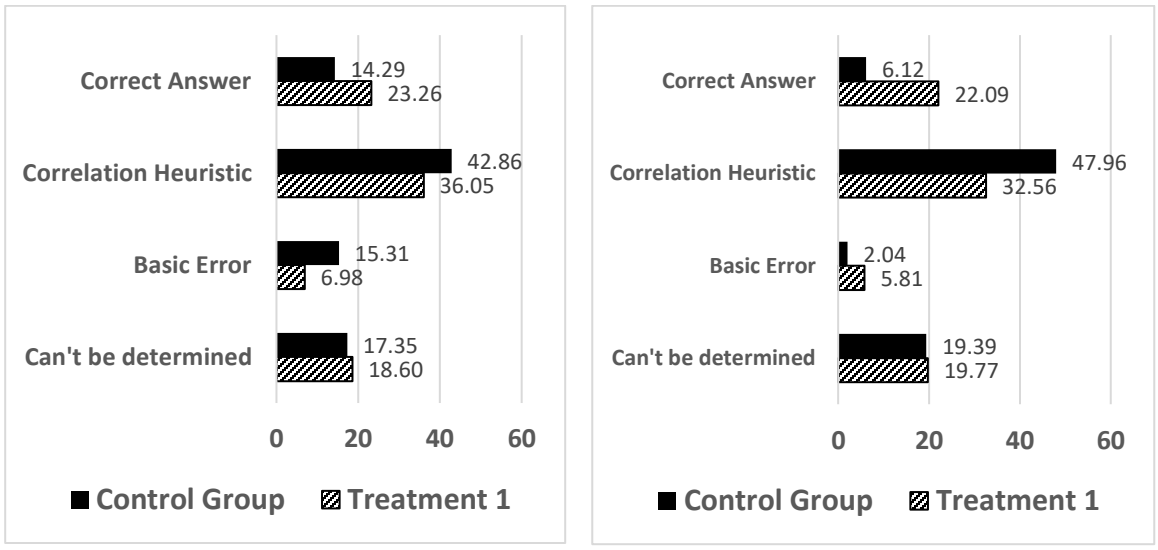
Heuristic error occurs when participants report minimum/maximum of net-flow as the minimum/maximum value of the stock. Basic Error occurs when participants report maximum of inflow as the maximum stock, and maximum of the outflow as the minimum of the stock. For example, if a participant answers the first stock question as 4 (wrong) and the second one as 21 (wrong), his or her Basic Error variable has the value of 2 and the variables Correct Answer, Correlation Heuristic, and CBD equal 0. If a participant answers the first stock question correctly and makes a basic error mistake in the other question, his or her Correct Answer and Basic Error variables will have the value of 1 while the Correlation Heuristic and CBD variables will be 0. These variables were then used to perform a formal regression analysis. Since our outcome variables are categorical variables with three levels, ordered logistic regression was our method of choice in this study.

Table 2. Coding description of participants' answers to question 3 (most people in the store) and question 4 (fewest people in the store).

	Most people in the store	**Fewest people in the store**
Correct Answer	$t \in \{12, 13, 14\}$	$t = 30$
Correlation Heuristic	$t = 8$	$t = 17$
Basic Error	$t = 4$	$t = 21$
CBD	"Can't be determined" was marked as the answer.	

Results

The results of Study 1 are shown in Figure 2. The ratio of correct answers to both stock questions has improved for participants in the treatment group. The ratio of correct answers to Question 3 has increased from 14.29% in the control group to 23.26% in the treatment group. The improvement is more considerable in Question 4 of the DS task: while 6.12% of participants in the control group answered Question 4 correctly, 22.09% of participants in the treatment group did so. The figure also shows the percentage of wrong answers due to the use of Correlation Heuristic, Basic Error, or selecting CBD. The correlation heuristic error is slightly lower for participants in the treatment group compared with the control group, while no trend can be seen for the other two types of error.



(a) (b)

Figure 2. Results of Study 1. (a) Question 3 of the original department store task: During which minute were the most people in the store? (b) Question 4 of the original department store task: During which minute were the fewest people in the store?

For a formal analysis of the results, we run six ordered logistic regression models. The results of these models are summarized in Table 3. Several variables are used in these models. The 'Treatment' variable, the main focus of our analysis, takes the value of 1 for participants in the treatment group and 0 for those in the control group. A significant coefficient for this variable shows that the effect of the intervention has been statistically significant. The first three models directly compare the difference between the control and treatment groups. In models 4–6, we include several independent variables to control for within-group variations. The 'Gender' variable equals 1 for females and 0 for males. The variable 'College' takes values of 1 for participants that have some level of education higher than high school and below a graduate degree, and the variable 'Graduate' is 1 for those with a graduate degree. These two variables are 0 for other participants. The variable math is the self-reported math expertise, which is an integer between 1 and 5.

Table 3. Summary of Ordered Logistic Regression Models for Study 1

	Models without demographic variables			**Models with demographic variables**		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6

Dependent Variables:	Correct Answer	Correlation Heuristic	Basic Error	Correct Answer	Correlation Heuristic	Basic Error
Independent variables						
Treatment	0.84*	-0.49†	-0.39	0.79*	-0.4	-0.34
	(0.35)	(0.28)	(0.2873)	(0.35)	(0.29)	(0.44)
Age				-0.01	0.01	-0.02
				(0.02)	(0.02)	(0.03)
Gender				-0.48	-0.23	0.6
				(0.37)	(0.29)	(0.43)
Math				-0.25	0.15	0.02
				(0.2)	(0.16)	(0.24)
College				0.7†	-0.5	-0.17
				(0.42)	(0.32)	(0.47)
Graduate				1.3*	-0.91†	-0.61
				(0.6)	(0.51)	(0.85)
χ^2 *(df)*	6.08 (1)	3.08 (1)	0.82 (1)	12.67 (6)	7.98 (6)	3.67 (6)
N	184	184	184	184	184	184

† $p < 0.1$ * $p < 0.05$ ** $p < 0.01$
Standard errors are presented inside parentheses.

Both models 1 and 4 show that the treatment intervention was successful in improving participants' performance in the SF task (odds ratio: 2.21). People in the treatment group answered more stock questions correctly compared with those in the control group. However, our intervention in this study does not have a significant effect on the Correlation Heuristics and Basic Error. Having a graduate degree had a positive effect on the participants' performance in the SF task. None of the other demographic variables had a significant effect on performance in the SF task. We tested the proportional odds assumption by using a likelihood-ratio test (the approximate likelihood ratio test of proportionality of odds was at $p>0.05$).[1]

Manipulation Check

[1] Using the omodel.ado package available at: http://econpapers.repec.org/software/bocbocode/s320901.htm

Measuring whether participants used their System 2 mode of thinking is not an easy procedure. In this study, following BG's original work, we look for specific indicators that can be signs of participants using their analytic mode of thinking. In the last step of our study, after participants answered the DS task, they were asked to explain how they answered the last two (i.e., stock) questions. Two researchers, who were not aware of the individual participant groups, coded their responses. The researchers were trained to use the following criteria for coding participants' responses: (1) recalling the trends of the graphs, (2) expressing a thinking strategy (rather than claiming they just needed to 'look' at the graph), and (3) acknowledging the difficulty of the task. These three criteria were used in BG's study and are claimed to be indicators of deeper thoughts (e.g., if they recall the graphs, they have probably thought more deeply about the questions).

Each of the researchers gave a '1' to the participants who show any of the abovementioned three indicators. For example, the answer "*More people entered than left up to and including the 13 minute mark. More people left than entered each minute after that.*" received a '1' in the 'recalling the trends' criteria. After the researchers coded participants' responses individually, they discussed the responses that their coding was at disagreement with each other. In some cases, the coders modified their response after this discussion. The Cohen's kappa coefficient is at $\kappa = 0.95$, which shows an almost perfect inter-rater agreement of coders (McHugh, 2012). We found that more participants in the treatment group showed traces of using the analytic mode of thinking in their justifications compared with those in the control group. While 29.1% of participants in the treatment group received at least one '1' from any of the coders, only 16.3% of participants in the control group received a '1' from our coders (p-value: 0.04).

STUDY 2

In Study 1, we replicated BG's results by demonstrating that asking participants analytical questions right before the DS task could improve their performance in the stock questions. While we observed the effects of using analytic mode of thinking, a deeper question remains: why does asking analytical questions lead participants to use their System 2 mode of thinking and better understand accumulation? It is also important to investigate whether other similar interventions can create a similar effect and improve performance in SF tasks.

We develop three hypotheses. First, asking for justification is known to be helpful in de-biasing people. Several studies have shown that when we try to find support for our decisions (process accountability), we are more likely to compare different alternatives and think more analytically about our decisions (for a review, refer to Lerner and Tetlock, 1999). For example, Fennema and Perkins (2008) report that asking for justification can be helpful in improving people's decision making when poor performance is due to low self-critical attention to the

task. So, the question is whether asking for justification helps improve participants' performances in the DS task by pushing them to think deeper about their answers. This is also a potential explanation for the first study; it might be the case that including analytical questions makes people search for justifiable solutions in the DS task too. The implication is that asking the participants to justify their answers while letting them change their answers to the SF task should have similar effects. We test this idea and call it the 'justification' intervention. Specifically, we conduct a test in which the analytical questions of Study 1 are replaced with questions that require participants to justify their answers to the DS task.

Second, it is possible that the first four questions of Study 1 required the participants to concentrate more since the questions were obviously non-trivial. When the DS task is asked immediately after those questions, it may imply that the DS task is also difficult. As a result, the participants expect the DS task to be difficult or tricky. If true, this means that the role of the first questions was to 'warn' the participants that there might also be something difficult or tricky with the DS task and it needs further consideration. Kahneman and Frederick (2002) explain how manipulation of attention can eliminate biases in cognitive tasks. Experimental studies have reported that warnings can reduce certain cognitive biases (Cheng and Wu, 2010; George et al., 2000; Block and Harper, 1991). Many other studies have claimed that warnings can influence people's behavior, making them more cautious about their tasks and environment, as long as the warnings are realistic and not too frequent (Ghaffarzadegan and Andersen 2012). The implication of this mechanism is that if one explicitly 'warns' the participants about the difficulty of the DS task instead of giving them the analytical questions, similar results should hold. We test this hypothesis in Study 2 by replacing the intervention from the first study with a simple warning message about the difficulty of the DS task, described in the method section. We call this the 'warning' intervention.

Third, we try to explicitly inform the participants about the two modes of thinking. Here we focus on providing information on analytic thinking, and how sometimes relying on 'fast' thinking can yield to mistakes. Overall, we may expect that if one reminds people of their different modes of thinking as well as their potential biases and asks them to use their analytic mode of thinking, the chances of finding correct answers in the DS task increase. Past studies suggest that providing explicit instructions to people can urge them to deliberately use their System 2 in problem solving (Evans, 2006). We test this hypothesis by replacing the intervention from the first study with a text that explains what 'analytic' and 'intuitive' modes of thinking are, and encourage them to take the analytic thinking approach in answering the DS task. We call this the 'explicit priming' intervention.

Obviously, multiple reasons may exist for the observed results in our first study, and any or none of the three new interventions may work. The purpose of the study is to shed more light on the potential reasons of the observed effect in the first study and to explore other methods that may help improve public understanding of accumulation. We selected these three

interventions as examples of different approaches to nudge people into using their analytic cognitive processes.

Method

As stated, here, we test three hypotheses. These hypotheses are operationalized as three interventions of treatment 2a (justification), treatment 2b (warning), and treatment 2c (explicit priming) as follows.

- Treatment 2a: Justification
  We ask the participants to justify their answers while letting them (and encouraging them to) modify their answers. We use the same justification questions that were used in the previous study. Everything is similar to the control group of Study 1, with the exception that they can modify their answers while explaining how they arrived at their answers.

- Treatment 2b: Warning
  We warn the participants about the difficulty of the stock questions. Everything in this experimental design is similar to the control group of Study 1, with one exception. In this group, right before the DS task, the following message is shown:

  > "Warning: In the next step, you will see a graph with four questions. The 3rd and the 4th questions are difficult. In our previous survey, more than 80% of the participants gave wrong answers to these questions. Also, in another survey at MIT, about half of the graduate students answered those questions incorrectly."

- Treatment 2c: Explicit Priming
  We inform the participants about the two modes of thinking. Everything in this design is similar to the control group of Study 1, with one exception: in this group, right before the DS task we explain the difference between the two modes of thinking (Appendix II). We then ask the participants to recall a situation ("like their investment decisions") when they used their analytic thinking skills and suppressed the temptation to rely only on fast thinking. They are then asked to write a paragraph about this experience and encouraged to follow a similar process in answering the SF task.[2]

[2] We acknowledge that the 'explicit priming' intervention in our study is not an ideal representation of an educational treatment. Education literature advocates the importance of experience and exercise in learning (Leron and Hazzan, 2009). Unfortunately, due to the online nature of our study we are not able to provide such a condition for our participants. Instead, we tried to include an example and asked for participants' past experiences in this intervention.

A total of 298 new participants (98 participants in Treatment 2a, 109 participants in Treatment 2b, and 91 participants in Treatment 2c) were involved in this study. Assignment of participants to treatment groups was random. Table 4 shows the self-reported demographics of the new 298 participants.

Table 4. Demographics of Participants Added to Study 2

Number of participants	298
Percentage of females	43.9%
Average age (SD)	34 (10.5)
High school education	31.2%
College-level education	61.7%
Graduate-level education	7.1%
Average level of self-reported mathematical expertise* (SD)	2.9 (0.9)

*: On a scale of 1–5; 2 to be equivalent to high school mathematics.

Results

Figure 3 shows the results of Study 2. We observe that changes in correct answers occur in the expected direction and that all treatments are successful in improving participants' performance, especially on the 4th question of the DS task (fewest people in the store). Statistically, however, only the warning treatment (Treatment 2b) significantly improves the results. The figure also shows the percentage of participants answering stock questions erroneously in the categories of Correlation Heuristic, Basic Error, and "Can't be determined." The warning treatment decreased the rate of the "Can't be determined" response considerably to both stock questions.

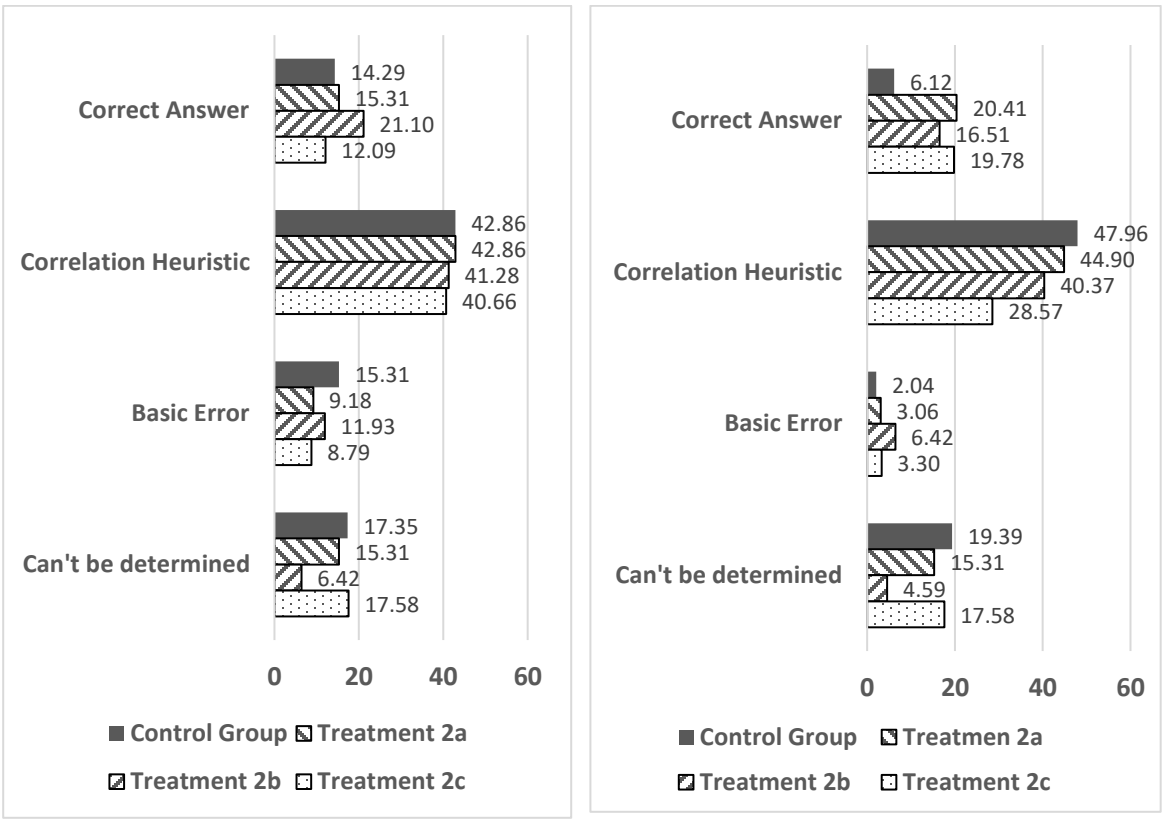


Figure 3. Results of Study 2. (a) Question 3 of the original department store task: During which minute were the most people in the store? (b) Question 4 of the original department store task: During which minute were the fewest people in the store?

Similar to the previous study, we used six ordered logistic regression models to analyze the results of Study 2. The results for these regressions are reported in Table 5. We observe that Treatment 2b (i.e., the warning intervention) is successful in increasing the correct answers to the stock questions (odds ratio: 2.05). Effects of other interventions are also in the desired direction, but are not statistically significant. Similar to Study 1, we tested the proportional odds assumption by using a likelihood-ratio test (the approximate likelihood ratio test of proportionality of odds was at $p>0.05$). The demographic variables Gender and Math are also significant in the Correct Answer model. As can be seen in Figure 3, there seems to be a difference between improvement in responses to Q3 and Q4. In particular, we see that all three interventions have increased the correct answers in Q4, while only the 'warning' treatment has increased the correct answers in Q3. From the analysis shown in Table 5, we conclude that only the warning intervention was successful in improving participants' performance. However, an analysis of the two questions separately (Appendix III) shows that all interventions were successful in significantly improving participants' performances in Q4.

Table 5. A Summary of Ordered Logistic Regression Models for Study 2.

	Models without demographic variables			**Models with demographic variables**		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Dependent Variables:	Correct Answer	Correlation Heuristic	Basic Error	Correct Answer	Correlation Heuristic	Basic Error
Independent Variables						
Treatment 2a	0.49	-0.06	-0.43	0.53	-0.05	-0.41
	(0.35)	(0.27)	(0.42)	(0.36)	(0.27)	(0.43)
Treatment 2b	0.67*	-0.18	-0.03	0.73*	-0.19	-0.05
	(0.33)	(0.26)	(0.38)	(0.34)	(0.26)	(0.39)
Treatment 2c	0.35	-0.49†	-0.36	0.37	-0.51†	-0.33
	(0.36)	(0.28)	(0.42)	(0.37)	(0.28)	(0.43)
Age				-0.01	0.01	-0.02
				(0.01)	(0.01)	(0.02)
Gender				-0.85**	0.12	0.5†
				(0.26)	(0.19)	(0.3)
Math				0.29*	-0.06	-0.08
				(0.15)	(0.11)	(0.17)
College				-0.01	-0.08	0.13
				(0.27)	(0.21)	(0.34)
Graduate				-0.86	0.26	0.42
				(0.61)	(0.41)	(0.59)
χ^2 *(df)*	4.42 (3)	3.74 (3)	1.67 (3)	25.7 (6)	7.22 (6)	6.56 (6)
N	396	396	396	396	396	396

† $p < 0.1$ * $p < 0.05$ ** $p < 0.01$

Standard errors are presented inside parentheses.

Manipulation Check

The procedure for manipulation check in this study is exactly the same as the one explained in Study 1. The Cohen's kappa coefficient is $\kappa = 0.97$, which shows an almost perfect inter-rater agreement of coders (McHugh, 2012). We expected all interventions to increase reliance on the analytic mode of thinking, and the results show that all treatments have increased the percentage of participants that show a trace of using their analytic modes of thinking from

16.3% in the control group to 29.6%, 26.6%, and 32.9% in Treatment groups 2a, 2b, and 2c, respectively ($p<0.1$).

DISCUSSION AND CONCLUSIONS

In this paper, we examined the problem of people's failure to understand the concepts of stocks and flows, and investigated methods of mitigating the problem. We built our theory of SF failure based on the dual-process theory from cognitive psychology (Kahneman and Frederick, 2002; Evans, 2003). Specifically, we postulated that the failure to correctly respond to an SF question is a result of the human tendency to use heuristic processes rather than analytic ones, and hypothesized that if people use their System 2 mode of thinking, they are more likely to answer the SF questions correctly. Our experimental studies show that priming people to use their analytic mode of thinking can modestly but significantly improve their answers to the stock questions in an SF task.

The study provides two major contributions to the literatures of complex systems and cognitive psychology. First, we replicate the results of the previous study by Baghaei Lakeh and Ghaffarzadegan (2015) in Study 1 and show that asking participants analytical questions right before the SF task improves their performance. Our results provide further evidence for the observed phenomenon. In contrast to BG's study, we used the original department store (DS) task and showed that the results of BG's study still holds. In sum, in contrast to many past studies that claim people do not understand complex systems, we claim that some of the observed failure in SF tasks is not necessarily related to people's lack of knowledge of accumulation, but rather about their cognitive mode of thinking; that is, if they think 'slower,'[3] they may understand accumulation.

Second, our study contributes to the literature of complex systems by showing that priming methods exist that may result in improving the understanding of basic concepts of complex systems. Specifically, we show that raising participants' attention by warning them about the difficulties of SF questions could improve their performance. These results are in line with the cognitive psychology literature showing that attention manipulation can be helpful in de-biasing people (Kahneman and Frederick, 2002). A potential reason for the effect of warning is that the study participants were invited to think more deeply about the question. We think that the warning message gave some participants the impression that the DS task was difficult and that they should focus. This led their System 2 thinking to more rigorously evaluate their answers and improved the rate of correct responses. As a result, the rate of correct answers for

[3] Here, we are using Kahneman and Frederick (2002) terminology.

participants who received warning increased from 14.29% to 21.10% in Question 3 and from 6.12% to 16.51% in Question 4.

In addition to these two major contributions, we obtain some encouraging results from two other interventions, which suggest future avenues of research. Specifically, we observe that asking participants to justify their answers (Treatment 2a) could improve participants' performance in one of the stock questions (Appendix III). This is in line with the argument that process accountability improves accuracy in cognitive tasks (Lerner and Tetlock, 1999). Interestingly, we observed that when people were asked to justify their answers, they were also more likely to change their answers. Furthermore, we tried to inform people to cautiously use their System 2 mode of thinking (Treatment 2c). We explained the concepts of the System 2 mode of thinking and asked them to use this system for the SF question. In this intervention, the participants relied less on correlation heuristic and their answers to at least one of the SF questions improved (Appendix III). We think that these results show how participants in this group became aware of their potential errors, which helped them to avoid the most common type of error in the accumulation tasks. However, the overall effect of the intervention on participants' performances was not statistically significant. As mentioned in Leron and Hazzan (2009), education should be accompanied by exercise and experience. One weakness of our 'explicit priming' treatment is that we could not provide such conditions perfectly. Further studies are needed to investigate the effect of informing people about their biases on their performances in SF tasks with a better design.

We find mixed results regarding the association between demographic variables and people's performance on the DS task. In Study 2, male participants performed significantly better than females. This is in agreement with prior studies (Veldhuis and Korzilius, 2016; Baghaei Lakeh and Ghaffarzadegan, 2015; Sterman, 2010) and we believe controlling for more variables eventually eliminates this effect. Moreover, having a graduate degree increased the chances of participants answering the stock questions correctly in Study 1. We did not observe this effect in Study 2. One potential reason for this variation might be due to the differences in participants' educational backgrounds. It is shown that engineering students are more likely to understand accumulation than management students (Kapmeier *et al.*, 2016). Further research is required to better understand the effect of demographic variables on people's understanding of accumulation.

Our study builds on several past SF failure studies by showing that many people fail to respond correctly to SF questions (Booth Sweeney and Sterman, 2000; Kapmeier, 2004; Cronin *et al.*, 2009; Sterman, 2010; Abdel-Hamid *et al.*, 2014; Qi and Gonzalez, 2015), but contrasts by questioning the assumption that the failure is only related to people's lack of knowledge about the concept of accumulation. We did not teach our participants about stocks and flows, but with simple interventions we were able to double the likelihood of answering the questions correctly. We agree that there might be specific challenges for implementing our

solution in real world cases. This study was focused on the laboratory setting and more investigations are needed before generalizing the solution for real world cases such as public understanding of global warming. This study is in line with a few others that have investigated various cognitive reasons that are responsible for people's poor performance on SF tasks (Veldhuis and Korzilius, 2016; Fischer and Gonzalez, 2016), but it also differs as it focuses on the dual-process modes of thinking.

This study has major limitations, and we invite readers to be cautious in interpreting the results. While these interventions were modestly successful in improving people's performance in SF tasks, the majority of the participants failed to respond to the questions correctly. More than 78% of participants in these studies answered at least one of the stock questions incorrectly. This shows that our interventions have a limited effect on people's performance in the DS task. In simple terms, our interventions, at best, influence only some people. While we stress the importance of being vigilant in mitigating SF failure, we would like to acknowledge that individuals' knowledge of SF can still play a critical role. Learning about the concept of accumulation might even become a more important factor as we move beyond simple laboratory experiments and look at real world tasks.

Moreover, the effect of interventions on mitigating correlation-heuristic error is in the expected direction but is not statistically significant (Table 5). More investigation is needed to understand the nature of the cognitive bias that causes this error. It remains unclear whether correlation heuristic error occurs due to participants being in their System 1 mode of thinking and using heuristics (cognitive shortcuts) or because they use their System 2 mode of thinking but employ the wrong analytic process. System 2 can suppress our tendency to blindly follow heuristics, but it cannot avoid systematic errors if the necessary tools for solving problems do not become available to us at the right time.

Finally, we should carefully consider the fact that our experiments were conducted on Amazon Mechanical Turk. The observed effects may not hold for populations with higher education levels, or for people with different cultural backgrounds. While our controlled study succeeds internal validity, generalization of the findings to daily-life tasks needs further investigations (Berinsky *et al.*, 2012).

An implication of our findings is that cognitive context matters. Answering accumulation questions is difficult for most people, but some may do a better job if they are primed to use their analytic mode of thinking. Our study shows that if we frame an SF task in a demanding cognitive context (i.e., by including the SF task in the context of other difficult analytical questions or by warning people about the difficulty of the SF task), we can increase the likelihood of correct answers.

Overall, this paper sheds more light on the effects of the analytic mode of thinking on performance in accumulation tasks. We claim that people's poor performance in accumulation tasks is related to how they think about these problems. We invite further research to better understand the underlying cognitive errors that lead to failure on SF tasks as well as in other complex systems-related concepts.

REFERENCES

Abdel-Hamid T, Ankel F, Batlle-Fisher M, *et al.* 2014. Public and health professionals' misconceptions about the dynamics of body weight gain/loss. *System Dynamics Review* **30**(1-2): 58-74.

Alter AL, Oppenheimer DM, Epley N, Eyre RN. 2007. Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General* **136**(4): 569-576.

Baghaei Lakeh A, Ghaffarzadegan N. 2015. Does analytical thinking improve understanding of accumulation? *System Dynamics Review* **31**(1-2): 46-65.

Berinsky AJ, Huber GA, Lenz GS. 2012. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis* **20**(3): 351-368.

Block RA, Harper DR. 1991. Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes* **49**(2): 188-207.

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

Buhrmester M, Kwang T, Gosling S. 2011. Amazon's Mechanical Turk a new source on inexpensive, yet high-quality, data? *Perspectives on Psychological Science* **6**(1): 3-5.

Cheng F, Wu C. 2010. Debiasing the framing effect: The effect of warning and involvement. *Decision Support Systems* **49**(3): 328-334.

Cronin MA, 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.

Darling-Hammond L. 2000. Teacher quality and student achievement. *Education policy analysis archives* **8**: 1.

Evans J. 2003. In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences* **7**(10): 454-459.

Evans J. 2006. The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychologival Bulletin & Review* **13**(3): 378-395.

Fennema MG, Perkins D. 2008. Mental budgeting versus marginal decision making: training, experience and justification effects on decisions involving sunk costs. *Journal of Behavioral Decision Making* **21**(3): 225-239.

Fischer H, Kapmeier F, Tabacaru M, Kopainsky B. 2015. The more you see the less you "get": On the importance of a higher-level perspective for understanding dynamic systems. In *Proceedings of the 2015 International System Dynamics Conference*. Available: http://www.systemdynamics.org/conferences/2015/proceed/papers/P1222.pdf

Fischer H, Gonzalez C. 2015. Making sense of dynamic systems: How our understanding of stocks and flows depends on a global perspective. Cognitive Science **40**(2): 496-512.

George JF, Duffy K, Ahuja M. 2000. Countering the anchoring and adjustment bias with decision support systems. *Decision Support Systems* **29**(2): 195-206.

Gervais WM, Norenzayan A. 2012. Analytic thinking promotes religious disbelief. *Science* **336**(6080): 493-496.

Ghaffarzadegan N, Anderson DF. 2012. Modeling behavioral complexities of warning issuance for domestic security: A simulation approach to develop public management theories. *International Public Management Journal* **15**(3): 337-363.

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

Hämäläinen RP, Luoma J, Saarinen E. 2013. On the importance of behavioral operational research: The case of understanding and communicating about dynamic systems. *European Journal of Operational Research* **228**(3): 623-634.

Kahneman D, Frederick S. 2002. Heuristics of intuitive judgment: extensions and applications. In Gilovich Th, Griffin D, Kahneman D. (ed.) *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press.

Kainz D, Ossimitz G. 2002. Can students learn stock-flow-thinking? An empirical investigation. In *Proceedings of the 2002 International System Dynamics Conference*. Available: http://www.systemdynamics.org/conferences/2002/proceed/papers/Kainz1.pdf

Kapmeier F. 2004. Findings from four years of bathtub dynamics at higher management education institutions in Stuttgart. In *Proceedings of the 2004 International System Dynamics Conference*. Available: http://www.systemdynamics.org/conferences/2004/SDS_2004/PAPERS/197KAPME.pdf

Kapmeier F, Happach RM, Tilebein M. 2016. Bathtub dynamics revisited: An examination of déformation professionelle in higher education. *Systems Research and Behavioral Science*. In-press.

Lerner JS, Tetlock PE. 1999. Accounting for the effects of accountability. *Psychological Bulletin* **125**(2): 255-275.

Leron U, Hazzan O. 2009. Intuitive vs analytical thinking: four perspectives. *Educational Studies in Mathematics* **71**(3): 263-278.

Louis MR, Sutton RI. 1991. Switching cognitive gears: From habits of mind to active thinking. *Human Relations* **44**(1): 55-76.

Mason W, Suri S. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods* **44**(1): 1-23.

McHugh ML. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* **22**(3): 276-282.

Pala O, Vennix J. 2005. Effect of system dynamics education on systems thinking inventory task performance. *System Dynamics Review* **21**(2): 147-172.

Paolacci G, Chandler J, Ipeirotis PG. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* **5**(5): 411-419.

Qi L, Gonzalez C. 2015. Mathematical knowledge is related to understanding stocks and flows: results from two nations. *System Dynamics Review* **31**(3): 97-114.

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

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

Tversky A, Kahneman D. 1974. Judgment under uncertainty: Heuristics and biases. *Science* **185**(4157): 1124-1131.

Veldhuis GA, and Korzilius H. 2016. Seeing with the mind: The relationship between spatial ability and inferring dynamic behaviour from graphs. *Systems Research and Behavioral Science*. In-press.

APPENDIX I

In this appendix the four analytical questions used in the intervention step of the treatment group in Study 1 is reported:

Question 1:

Estelle states: When I went fishing the other day, every fish that I caught was a salmon, and every salmon I saw I caught.

Of the following statements listed below, which one can be concluded from the observations of Estelle?

A. Salmon was the only fish that Estelle saw while she was fishing.

B. While Estelle was fishing, she caught no fish other than salmon.

C. In the area that Estelle fished, there were no other fish.

D. All of the fish that Estelle saw she caught.

E. Estelle did not see any other fish while she was fishing.

Source: www.testprepreview.com

Question 2:

Please look at the figures below carefully. Does the figure on the left include the figure on the right?

Source: Witkin HA. 1950. Individual differences in ease of perception of embedded figures. *Journal of Personality* **19**(1): 1-15.

Question 3:

The staff of a grocery store have been surveyed and the following information is available:

- 16 people were wearing black shoes

- 20 people were wearing white shirts

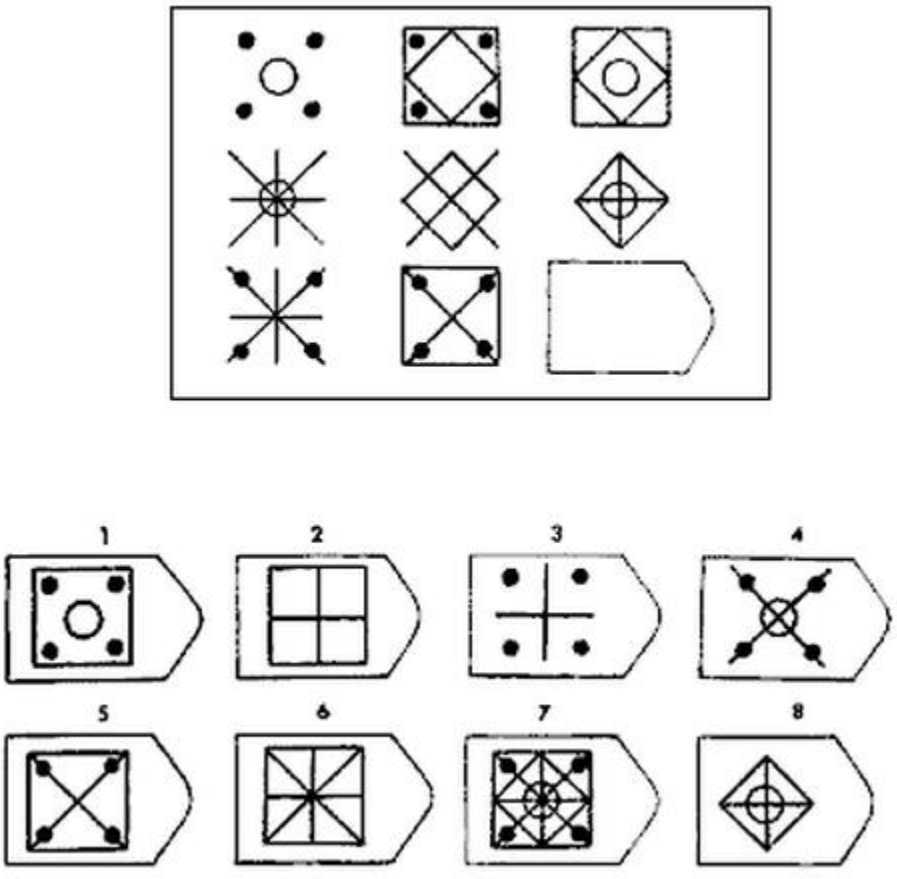
- 12 people were wearing both black shoes and white shirts

- 4 people were wearing neither black shoes nor white shirts

How many people were working in the store?

Question 4:

Which of the numbered diagrams at the bottom would complete the arrangement at the top?



Source: www.raventest.net

APPENDIX II

In this appendix, we are presenting the text which was used as intervention in the "educational" treatment of the second Study:

Many scholars including the noble prize winner Daniel Kahneman believe that people possess two modes of thinking. One mode operates automatically, quickly, and with little or no effort. They call this mode of thinking "fast." The other mode of thinking is deliberate, slow, and effortful. This mode of thinking is called "analytical." To see how these two modes of thinking work, let's consider this question (: "A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?"… Do you think the answer is $0.10? No! That is the answer people quickly come up with while they are not using their analytical mode of thinking. But if one uses analytical thinking, he/she would say if the cost of a bat is X and the cost of a ball is Y, then X+Y=$1.10, and X-Y=$1.0. Then, adding both sides of the equations, he/she will find that

(X+Y)+(X-Y)= $1.10+$1.0

2X=$2.10

X=$1. 05, and Y=$0.05. Thus the ball costs $0.05!

This is an example of the difference between fast thinking and analytical thinking. Analytical thinking requires slow and careful effort such as what we do when we solve a math problem.

Now think about a situation when you used your analytical thinking skills and suppressed the temptation to rely only on your fast thinking. Maybe when buying something that was on a special deal with some percentage discount, or managing your costs, or deciding to get a loan, or calculating calories before eating a high-calorie cake, or other similar situations when after an analytical evaluation you found something different than your initial quick judgment. Write a paragraph about your experience.

In the next step, you will see a graph with 4 questions. The 3rd and the 4th questions are difficult and need analytical thinking. Please use your analytical thinking skills to answer those questions.

Appendix III

In this appendix, we report the results of two additional logistic regression models for each of the stock questions (questions 3 and 4) in Study 2. As Figure 3 of the paper showed, the effects of the treatments in Study 2 seem to differ for questions 3 and 4. Here, we analyze each question separately.

As Table A1 shows none of our treatments were successful in significantly increasing the ratio of correct answers for question 3, but as Table A2 depicts all of our treatments were successful in increasing the ratio of correct answers for question 4. Moreover, Treatment 2c (the "educational" treatment) has decreased the ratio of correlation heuristic in question 4 significantly.

Table A1: Summary of logistic regression models for question 3 of study 2

	Models without demographic variables			**Models with demographic variables**		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Dependent Variables:	Correct Answer	Correlation Heuristic	Basic Error	Correct Answer	Correlation Heuristic	Basic Error
Independent variables						
Treatment 2a	0.08	0.00	-0.58	0.12	-0.001	-0.58
	(0.4)	(0.29)	(0.45)	(0.41)	(0.29)	(0.45)
Treatment 2b	0.43	-0.06	-0.29	0.56	-0.06	-0.3
	(0.37)	(0.28)	(0.41)	(0.38)	(0.29)	(0.42)
Treatment 2c	-0.19	-0.09	-0.63	-0.18	-0.08	-0.62
	(0.43)	(0.29)	(0.46)	(0.44)	(0.3)	(0.47)
Age				-0.01	0.01	-0.02
				(0.01)	(0.01)	(0.02)
Gender				-0.59*	-0.17	0.47

				(0.3)	(0.21)	(0.33)
Math				0.25	-0.14	-0.12
				(0.17)	(0.12)	(0.19)
College				-0.26	-0.1	0.13
				(0.32)	(0.23)	(0.36)
Graduate				-0.24	0.58	-0.6
				(0.58)	(0.4)	(0.77)
χ^2 *(df)*	3.32 (3)	0.15 (3)	1.67 (3)	11.9 (8)	5.26 (8)	7.72 (8)
N	396	396	396	396	396	396

† p < 0.1 * p < 0.05 ** p < 0.01

Standard errors are presented inside parentheses.

Table A2: Summary of logistic regression models for question 4 of study 2

	Models without demographic variables			**Models with demographic variables**		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Dependent Variables:	Correct Answer	Correlation Heuristic	Basic Error	Correct Answer	Correlation Heuristic	Basic Error
Independent variables						
Treatment 2a	1.37**	-0.12	0.42	1.41**	-0.14	0.53
	(0.49)	(0.29)	(0.92)	(0.5)	(0.29)	(0.94)
Treatment 2b	1.11*	-0.31	1.19	1.12*	-0.34	1.27
	(0.49)	(0.28)	(0.81)	(0.51)	(0.29)	(0.83)
Treatment 2c	1.33**	-0.83**	0.49	1.3*	-0.88**	0.59
	(0.5)	(0.31)	(0.92)	(0.51)	(0.31)	(0.94)

Age				-0.002	0.01	-0.001
				(0.01)	(0.01)	(0.03)
Gender				-1.13**	0.39†	0.08
				(0.34)	(0.21)	(0.56)
Math				0.44*	0.01	0.06
				(0.18)	(0.12)	(0.32)
College				0.11	-0.05	-0.15
				(0.34)	(0.24)	(0.67)
Graduate				-1.29	0.05	1.64
				(0.79)	(0.4)	(0.68)
χ^2 *(df)*	11.15 (3)	8.61 (3)	2.92 (3)	37.47 (8)	13 (8)	8.35 (8)
N	396	396	396	396	396	396

† $p < 0.1$ * $p < 0.05$ ** $p < 0.01$

Standard errors are presented inside parentheses.