NOTES AND INSIGHTS

Graphical features of flow behavior and the stock and flow failure

Cleotilde Gonzalez,[a]* Liang Qi,[b] Nalyn Sriwattanakomen[a] and Jeffrey Chrabaszcz[a]

Syst. Dyn. Rev. **33**, 59–70 (2017)

Introduction

All accumulation problems are ruled by the same principles. A stock increases when the inflow is greater than the outflow, decreases when the outflow is greater than the inflow, and is stable when the inflow and the outflow are equal over a time period. For example, to maintain our weight, we must burn as many calories as we consume; to lose weight, we must burn more calories than we consume; and to gain weight, we must consume more calories than we burn. This reasoning seems almost too simple to be a problem for study, but research has shown exactly the opposite. In the past decade, evidence has mounted demonstrating our inability to understand these basic concepts of accumulation (Abdel-Hamid *et al.*, 2014; Brunstein *et al.*, 2010; Cronin and Gonzalez, 2007; Cronin *et al.*, 2009; Gonzalez and Wong, 2012; Sweeney and Sterman, 2000). This problem, called *stock–flow* (*SF*) *failure*, is robust and difficult to overcome (Cronin *et al.*, 2009). People fail to integrate the information about the inflow and outflow over time and consequently make erroneous judgments about the levels of a stock at different points in time. Researchers have found that people often use a *correlation heuristic* (Cronin *et al.*, 2009) that leads them to make judgments about the level of a stock based solely on the pattern of the flows.

A number of researchers have questioned whether SF failure results from a difficulty interpreting graphs of behavior over time or from the type of graphical representation used in these problems. For example, Cronin and Gonzalez (2007) suggested that the type of representation of dynamic systems is a critical source of errors in judging the relationship between flows and the stock. Cronin *et al.* (2009) varied the graphical pattern of behavior over time of inflows and outflows, and concluded that more complex patterns between inflows and outflows (e.g., an inverted U-shape and a line versus two parallel lines) result in less accurate judgments of accumulation and more reliance on the correlation heuristic. In their study, however, the variations in the patterns

[a] Dynamic Decision Making Laboratory, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, U.S.A
[b] Health Service, Second Military Medical University, Shanghai, China
* Correspondence to: Cleotilde Gonzalez, Dynamic Decision Making Laboratory, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, U.S.A. E-mail: coty@cmu.edu

Accepted by Ignacio Martinez-Moyano, Received 30 September 2016; *Revised 7 March 2017*; *Accepted 13 March 2017*

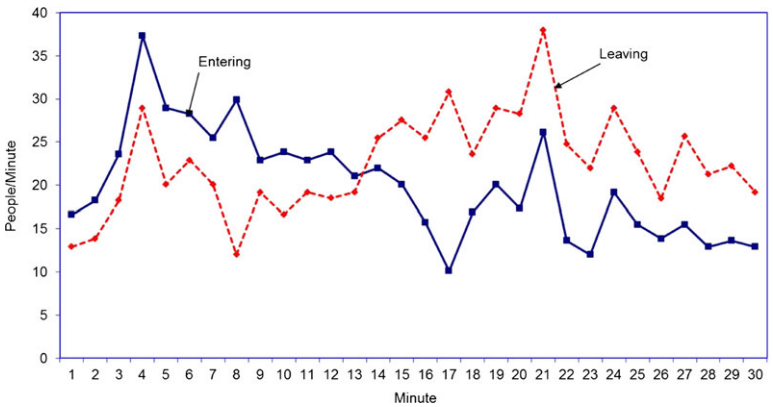
System Dynamics Review
System Dynamics Review vol 33, No 1 (January/March 2017): 59–70
Published online 28 May 2017 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/sdr.1570

of behavior of the flows also changed the resulting pattern of stock behavior over time (i.e., the shape of the stock line). These changes make it difficult to conclude whether the increased difficulty results from the more sophisticated graphical patterns of inflow and outflow or from the more complex pattern of the resulting stock. The present research provides additional insights into SF failure by demonstrating that when the resulting pattern of stock behavior over time is kept constant the graphical features of the patterns of inflow and outflow do not relate to the SF failure.

We used a common simple stock-and-flow task: the "department store" (DS) task, illustrated in Figure 1 (Sterman, 2002). A graph shows the number of people entering and leaving a department store each minute over a 30-minute interval, and participants are asked four questions. Two questions of interest (questions 3 and 4) test whether participants can infer the stock's behavior

Fig. 1. The department store (DS) task. [Colour figure can be viewed at wileyonlinelibrary.com]

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

DOI: 10.1002/sdr

from the behavior of the flows (Cronin and Gonzalez, 2007; Cronin *et al.*, 2009; Sterman, 2002). Evidence for SF failure is based on the low percentage of correct responses to these questions.

To answer questions 3 and 4 correctly, one needs to observe the relationship between the inflow and outflow, their crossing point, and the size of the areas before and after the two lines cross (Sterman, 2002). People should recognize that more people leave than enter the department store for a longer time period after the crossing point; therefore the minute at which there are the most people in the store is the crossing point (t = 14) and the minute where there are the fewest people in the store is the last minute (t = 30). People should be able to visually perceive the crossing point of the flows to discriminate the size of the areas between the curves and to determine which area is bigger (in Figure 1 the area after the crossing point is twice the size of the area before the crossing point, 1:2).

Fischer and Gonzalez (2016) suggest that the graphical elements of the DS task may influence the way humans process information (i.e., whether they look at the "forest" or the "trees"). Research by Korzilius et al. (2014), in which participants were asked to think aloud while solving the DS task, suggests that most people do focus on the local elements: participants concentrated on the peaks and the difference between the peaks. Relatedly, Fischer and Gonzalez (2016) found that global priming decreased SF failure relative to local priming. Presumably, global priming led participants to focus on the gestalt of the DS task—that is, the areas between the flow curves—rather than on the local elements (i.e., the peaks in the curves).

The present study tests how the features of graphical patterns of flow behavior over time may influence SF success. If more attention to the area between the curves suggests global processing, it follows that increasing the salience of the difference in the areas between the curves and the ease of discrimination of graphical patterns would increase SF success. Importantly, in this study, the behavior over time of the resulting stock in the DS task (initially increasing to a maximum point, and then decreasing) was the same in all conditions; we only manipulated the graphical representations of the inflow and outflow behavior over time and the salience of the difference in the areas between the flow curves. Specifically, we manipulated the point of the crossing of the flows, the scale of the areas between the flows, and the location of the bigger area between the flows.

Method

Participants

There were 957 participants. Their mean age was 28 years (standard deviation = 12.08), and 66.04 percent were female. The median education

DOI: 10.1002/sdr

level of participants was "some college" and the mean education level was a 2-year college degree. All participants were recruited online through Amazon Mechanical Turk and paid $0.50 for participation.

Procedure

After completing a consent form, participants performed a *visual judgment* accuracy task. In this part of the experiment, our aim was to determine which combinations of graphical elements led to better discernment of which area between the two flow curves was larger. To this end, we created 18 versions of the original DS graph (see Figure 2) by manipulating three graphical elements: placement of the *crossing* point of the flows (left, middle, or right side of the graph); *scale* of the bigger area (two, three, or four times as big); and *location* of the bigger area (before or after the crossing point). Crossing was coded as L (left, t = 10), M (middle, t = 15), or R (right, t = 20). Scale was coded as ratios: 1:2, 1:3, and 1:4. Finally, the location of the bigger area was indicated by the order of the numbers in the ratios. For example, R1:2 denotes a graph whose right area (i.e., the area after the crossing) is twice as big as the left and whose crossing occurs on the right side of the graph. In contrast, R2:1 denotes a graph whose left area (i.e., before the crossing on the right side of the graph) is twice as big as the right. In the visual judgment task, participants were asked to indicate as quickly as possible which area of the graph was bigger (left or right). They were asked to judge the size of the area for all 18 versions in random order.

Next, to determine whether visual judgment accuracy predicted success on the SF questions, we gave participants one of the 18 versions of the DS graph selected at random and asked them to solve the DS task with the four SF questions. For each question, participants typed in a number indicating a minute in the graph or selected the "Can't be determined" box if they believed the information could not be determined from the graph. Responses to the SF questions were coded as correct if they were between −1 minute and +1 minute away from the correct answer.

Results

The results answer four questions. First, how does the manipulation of crossing, scale, and location affect visual judgment accuracy (measured dichotomously: correct or incorrect) on the DS graph? Second, what effect does the manipulation of the patterns of behavior-over-time graphs have on SF accuracy (measured dichotomously)? Third, does visual judgment accuracy predict SF accuracy? Fourth, what types of SF judgment errors do participants

DOI: 10.1002/sdr

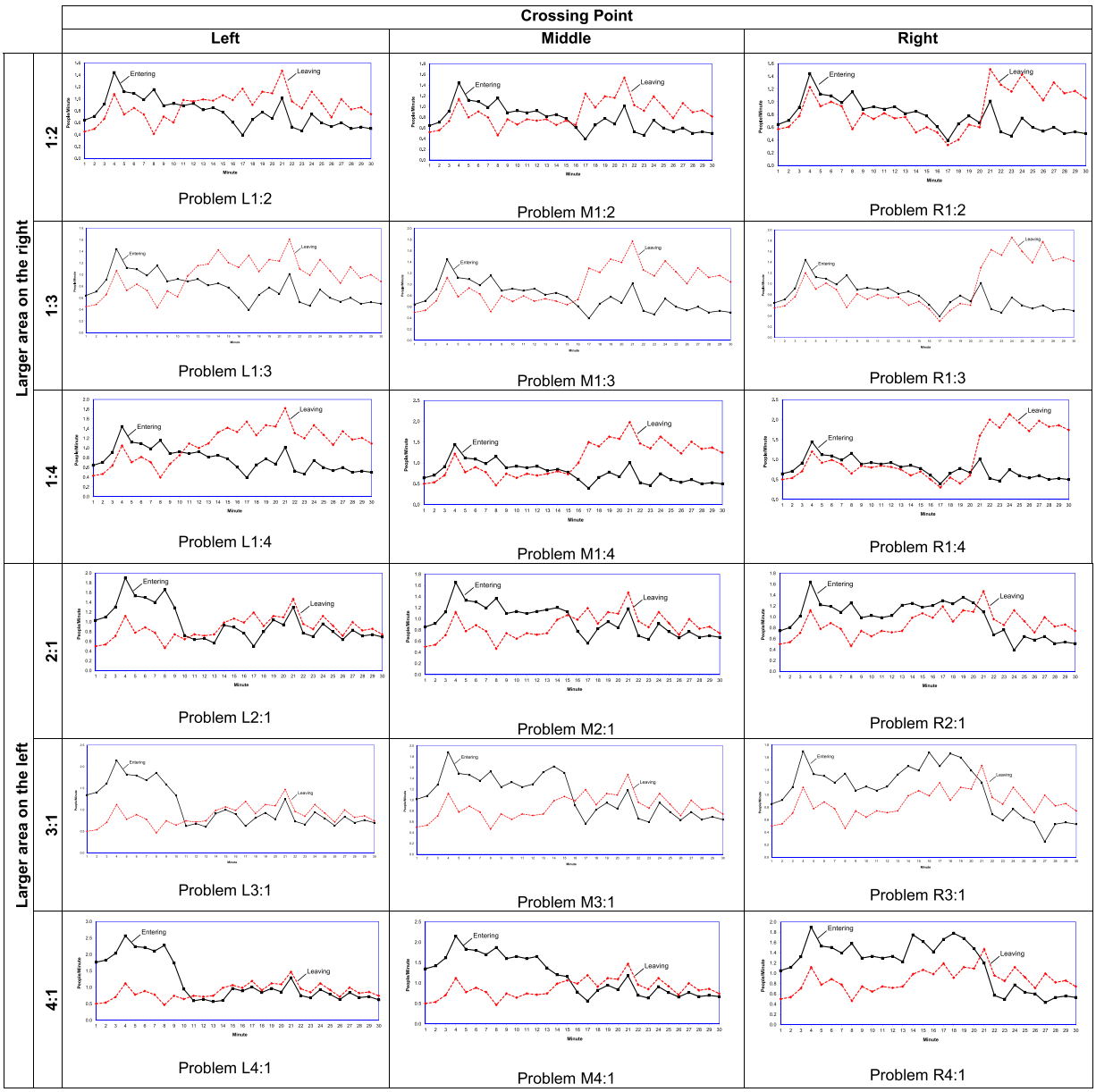


Fig. 2. Eighteen versions of the DS task graph used in the experiment. The solid lines represent people entering and the dotted lines represent people leaving over the course of 30 minutes (x axis). The y axis denotes the total number of people per minute. [Colour figure can be viewed at wileyonlinelibrary.com]

DOI: 10.1002/sdr

make? All analyses were conducted using the R statistical computing language (R Core Team, 2016) and the lme4 package (Bates *et al.*, 2015).

Visual judgment accuracy

Overall, participants were very accurate on the visual judgment task. The 95 percent confidence interval (CI) for success rate across all conditions was between 97.7 percent and 98.4 percent. (All ranges are 95 percent CIs unless otherwise noted.) We built a multilevel logistic regression using graphical elements (crossing, scale, and location) to predict visual judgment accuracy. This model included varying intercepts by participant to account for the dependence between repeated observations from each participant (i.e., the 18 trials) (Gelman and Hill, 2006). Perfect accuracy for all 18 responses in roughly 59 percent of participants prevented us from fitting any interaction terms. As shown in Table 1 and Figure 3, the model predicts significantly higher accuracy when the crossing point was on the right or in the middle (as opposed to on the left); when the larger area was three or four (as opposed to two) times larger than the smaller area; and when the location of the larger area was before the crossing (as opposed to after).

Stock–flow accuracy

The group accuracy rates for each of the four SF questions were calculated by dividing the number of correct responses by the total number of participants. These rates and their frequencies can be found in Table 2. The vast majority of participants responded correctly to Q1 (*95.9–96.5 percent*) and Q2 (*92.3–93.1 percent*), and overall accuracy was very poor for Q3 (*17.6–18.8 percent*) and

Table 1. Coefficients for the multilevel model predicting visual judgment accuracy using the graphical manipulations and varying intercept by participant. Intercept corresponds to the condition in which the crossing point is on the left of the display, the larger area is twice as big as the smaller area, and the larger area is after the crossing point (L1:2 in Figure 2)

Fixed effects				
	Coefficient	Coefficient SE	*z*-value	*p*-value
Intercept (L1:2)	3.09	0.12	25.60	$<2 \times 10^{-16}$
Crossing (M)	0.87	0.094	9.24	$<2 \times 10^{-16}$
Crossing (R)	0.30	0.084	3.59	0.0003
Scale (1:3)	0.65	0.086	7.57	4×10^{-14}
Scale (1:4)	1.00	0.093	10.75	$<2 \times 10^{-16}$
Location (before)	0.16	0.073	2.19	0.028
Random effects				
	Variance			
Participant	3.219			
Residual	1			

DOI: 10.1002/sdr

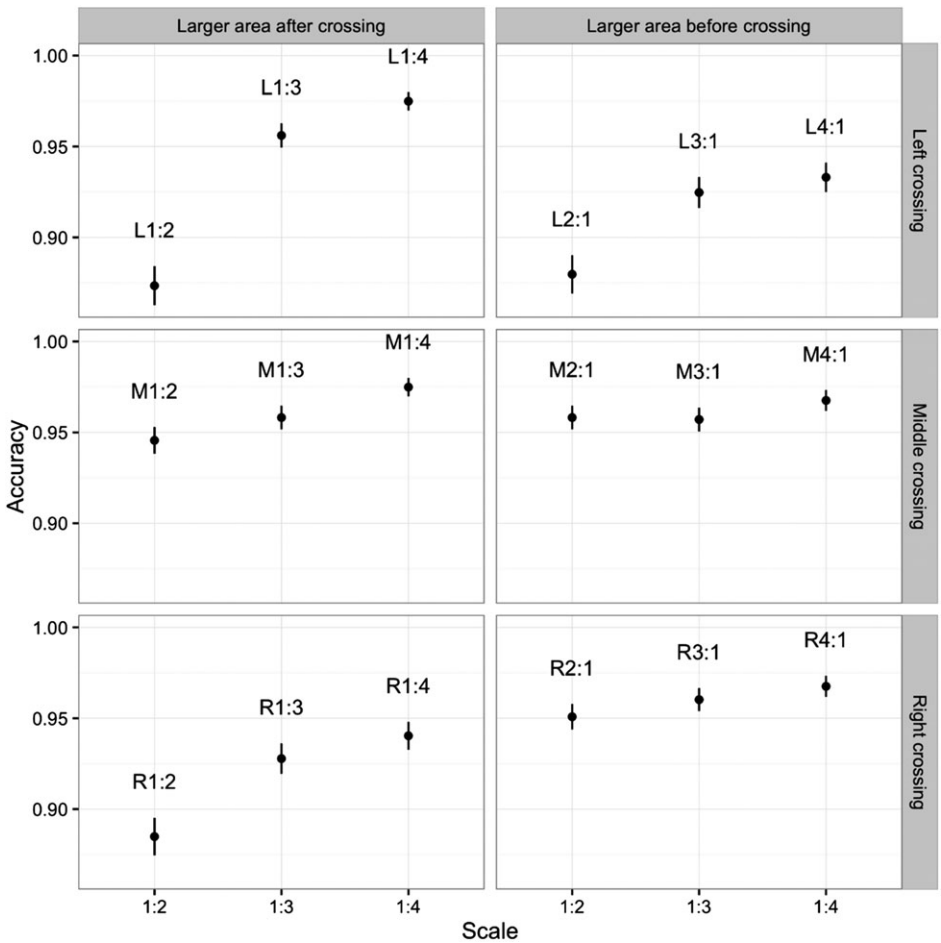


Fig. 3. Average accuracy rates in each condition of the visual judgment task. The horizontal axis corresponds to the scale factor, while the horizontal facets correspond to the location factor and the vertical facets correspond to the crossing factor

Table 2. Distributions of correct responses to the four questions in the DS task

	Question			
	Q1: Most entering	Q2: Most leaving	Q3: Most in store?	Q4: Fewest in store?
Correct responses	920	886	174	124
Frequency	96.23%	92.68%	18.20%	12.97%

Q4 (*12.5–13.5 percent*). The scores of the MTurk workers for Q3 and Q4 were lower than any of those reported in past research (Sterman, 2002; Cronin *et al.*, 2009; Cronin and Gonzalez, 2007), but they are similar to the rates found in other studies conducted on MTurk recently (Fischer and Gonzalez, 2016;

DOI: 10.1002/sdr

Qi and Gonzalez, 2015). As discussed in Qi and Gonzalez (2015), these differences may be due to the variances in settings and methods, as well as the level of education of the participants involved.

We used a logistic regression to assess whether graphical elements predicted accuracy on Q3 and Q4 from Table 2. Each participant answered each of these questions only once with regard to a particular combination of graphical elements, so no multilevel model was necessary. For each of these models, we used crossing, scale, location, and their two-way and three-way interactions to predict SF question accuracy. None of these predictors reached significance.

Effect of visual judgment accuracy on stock–flow accuracy

We built another regression model to test the effect of visual judgment accuracy on SF question (Q3 and Q4) accuracy. We scored visual judgment accuracy in two ways. The multilevel model (MLM) method used the individual intercept estimates from the model reported in Table 1 as an indicator of accuracy. For robustness, we also ran similar models with individual accuracy calculated as the proportion of correct visual judgments. Because the SF questions were associated with different arrangements of the DS graph, we again used multilevel logistic models to control for the experimental effects of crossing, scale, and location. Finally, we calculated Bayes factors on the relationship between visual judgment accuracy and SF question accuracy to assess the relative support for or against the predictive value of visual judgment accuracy for SF question accuracy (Rouder *et al.*, 2009). Contrary to predictions, as shown in Table 3, neither the multilevel models nor the Bayes factors supported a relationship between visual judgment accuracy and accuracy on either Q3 or Q4. In fact, the Bayes factors indicated that regardless of the method of scoring visual accuracy the data

Table 3. Slope coefficients from the multilevel models predicting Q3 and Q4 accuracy using either the intercept estimates from the model in Table 1 (MLM) or the proportion of correct responses on the visual judgment task [p(Cor)]. Bayes factors are presented as relative support for the null hypothesis of a coefficient value of zero

Question	Accuracy calculation method	Coefficient	Coefficient SE	*z*-value	*p*-value	Bayes factor
3	MLM	0.012	0.026	0.48	0.63	19.00
	p(Cor)	0.22	0.28	0.78	0.44	24.59
4	MLM	0.026	0.025	1.01	0.31	27.82
	p(Cor)	0.38	0.32	1.20	0.23	28.58

DOI: 10.1002/sdr

strongly favored the null hypothesis of no effect of visual judgment accuracy on SF accuracy (Jeffreys, 1998; Kass and Raftery, 1995).

SF error types

Incorrect responses were coded into different error types according to a scheme from previous research (Cronin and Gonzalez, 2007; Cronin *et al.*, 2009, Qi and Gonzalez, 2015). The rates for each error type were calculated by dividing the number of times the error was committed for each question by the total number of errors ($n = 1720$). The distribution of errors is shown in Table 4. In agreement with past research, common errors in Q3 and Q4 were indicative of the correlation heuristic. Participants answered with the time of *peak inflow* (5.99 percent) or the time of *peak net inflow* (17.44 percent) in Q3. They answered with the time of *peak outflow* (2.44 percent) or the time of *peak net outflow* (12.15 percent) in Q4. Thus our findings demonstrate a modestly high usage of the correlation heuristic.

Discussion

While the design of the graphical patterns of flow behavior over time determined the visual discriminability of the areas, we found no evidence of

Table 4. Distributions of error types in the DS task for each question

Error type	Q1	Q2	Q3	Q4
	0	1	**103**	5
Peak inflow	0.00%	0.06%	**5.99%**	0.29%
	0	16	**300**	22
Peak net inflow	0.00%	0.93%	**17.44%**	1.28%
	1	0	9	**42**
Peak outflow	0.06%	0.00%	0.52%	**2.44%**
	0	9	9	**209**
Peak net outflow	0.00%	0.52%	0.52%	**12.15%**
	2	3	3	28
Start point ($t = 1$)	0.12%	0.17%	0.17%	1.63%
	0	1	5	94
End point ($t = 30$)	0.00%	0.06%	0.29%	5.47%
	0	3	0	12
Crossing point	0.00%	0.17%	0.00%	0.70%
	5	5	255	270
Cannot be determined	0.29%	0.29%	14.83%	15.70%
	28	32	98	150
Other	1.63%	1.86%	5.70%	8.72%
	36	**70**	**782**	**832**
Total	**2.09%**	**4.07%**	**45.47%**	**48.37%**

DOI: 10.1002/sdr

a relationship between the accuracy of visual judgment in these graphs and success on judging the stock behavior in Q3 and Q4.

The areas in the DS graph were easier to discriminate when the crossing point was on the right, and when the larger area on the left was at least three times larger than the smaller area on the right (after the crossing point). Examples of easier graphs are R3:1 and R4:1; other graphs had area sizes that were more difficult for people to discriminate. These results generally agree with findings on visual discrimination from graphs (e.g., Treisman and Gelade, 1980; Wickens, 1992). However, our results strongly suggested *no effect* of visual judgment accuracy on SF success. Participants were near-perfect in their visual judgment accuracy but exhibited extremely poor performance on the SF accumulation questions. Thus the patterns of flow behavior over time do *not* relate to SF failure, suggesting that the difficulty in judging accumulation may be rooted in the pattern of stock behavior over time itself.

An analysis of the errors that participants committed on the SF accumulation questions suggests a moderate incidence of the correlation heuristic. Specifically, the frequencies of *peak inflow* and *peak net inflow* errors for Q3, and *peak outflow* and *peak net outflow* errors for Q4 suggest that the correlation heuristic contributed to SF failure. However, these incidences were lower than those reported in prior studies with a similar population (e.g., Qi and Gonzalez, 2015), such as 59.1 percent for Q3 and 44.7 percent for. Moreover, the "cannot be determined" error occurred more frequently for Q4 than in these prior studies. These errors may occur due to a general tendency to concentrate on the details of the graph rather than on the gestalt: being "local" rather than "global" processors (Fischer and Gonzalez, 2016; Korzilius *et al.*, 2014). The lower error rate compared to prior studies may be due to a possible priming of "global" processing caused by the visual judgment task presented before the DS task. It is also possible that in this study, given the design of the behavior-over-time graphs, participants could not always rely on local elements (i.e., the peaks in the curves) as in previous studies, and they felt more compelled to check the "cannot be determined" box.

While we conclude that the graphical features of the behavior of flows over time in the DS task do not relate to SF failure, there are open questions that require further empirical investigation. For example, Fischer and Gonzalez (2016) found a lower rate of SF failure when Q3 and Q4 were worded to highlight the global trend of the flows rather than the local, one-point-in-time aspects. The difficulty originating from the format of the questions needs to be investigated further through experimentation.

Biographies

Cleotilde Gonzalez is a research professor in the department of Social and Decision Sciences and the founding Director of the Dynamic Decision Making Laboratory at Carnegie Mellon University. Her research focuses on

DOI: 10.1002/sdr

behavioral studies of human dynamic choice and computational representations of cognitive processes of dynamic decision making.

Liang Qi is a researcher sponsored by Shanghai Pujiang Program and a lecturer in Department of Health Service at Second Military Medical University, Shanghai, China. His research focuses on dynamic system cognition, system dynamics, and judgment and decision making.

Nalyn Sriwattanakomen is a research associate at the Dynamic Decision Making Laboratory at Carnegie Mellon University.

Jeffrey Chrabaszcz is a postdoctoral researcher at the Dynamic Decision Making Laboratory at Carnegie Mellon University. His research focuses on Bayesian modeling of human memory, judgment, and choice.

References

Abdel-Hamid T, Ankel F, Battle-Fisher M, Gibson B, Gonzalez-Parra G, Jalali M, Kaipainen K, Kalupahana N, Karanfil O, Marathe A, Martinson B. 2014. Public and health professionals' misconceptions about the dynamics of body weight gain/loss. *System Dynamics Review* **30**(1–2): 58–74. https://doi.org/10.1002/sdr.1517

Bates D, Maechler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**(1): 1–48.

Brunstein A, Gonzalez C, Kanter S. 2010. Effect of domain experience in the stock-flow failure. *System Dynamics Review* **26**(4): 347–354. https://doi.org/10.1002/sdr.448

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.

Fischer H, Gonzalez C. 2016. Making sense of dynamic systems: How our understanding of stocks and flows depends on a global perspective. *Cognitive Science* **40**: 1–17.

Gelman A, Hill J. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press: Cambridge, U.K.

Gonzalez C, Wong H. 2012. Understanding stocks and flows through analogy. *System Dynamics Review* **28**(1): 3–27. https://doi.org/10.1002/sdr.470

Jeffreys H. 1998. *The Theory of Probability*. Oxford University Press: Oxford.

Kass RE, Raftery AE. 1995. Bayes factors. *Journal of the American Statistical Association* **90**(430): 773–795.

Korzilius H, Raaijmakers S, Rouwette E, Vennix J. 2014. Thinking aloud while solving a stock–flow task: surfacing the correlation heuristic and other reasoning patterns. *Systems Research and Behavioral Science* **31**(2): 268–279.

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.

DOI: 10.1002/sdr

R Core Team. 2016. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria Available: https://www.R-project.org/.

Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. 2009. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review* **16**(2): 225–237. https://doi.org/10.3758/PBR.16.2.225

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

Sweeney LB, Sterman JD. 2000. Bathtub dynamics: Initial results of a systems thinking inventory. *System Dynamics Review* **16**(4): 249–286. https://doi.org/10.1002/sdr.198

Treisman AM, Gelade G. 1980. A feature-integration theory of attention. *Cognitive Psychology* **12**(1): 97–136.

Wickens CD. 1992. *Engineering Psychology and Human Performance*. (2nd ed.) HarperCollins Publishers Inc.: New York.

DOI: 10.1002/sdr