System Dynamics Review
System Dynamics Review (2010)
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/sdr.448

NOTES AND INSIGHTS

Effects of domain experience in the stock–flow failure

Angela Brunstein,[a] Cleotilde Gonzalez[a]* and Steven Kanter[b†]

Abstract

Misperceptions of stock and flow relationships are pervasive and an important problem to solve in system dynamics. Prior studies have shown that individuals perform poorly on accumulation problems, even when considering relatively simple systems, an effect termed the *Stock–Flow (SF) failure*. This study examines the effects of domain experience in overcoming the SF failure. We compared performance of medical students and undergraduates with no medical education on accumulation problems in medical and general domains. Medical students performed better than undergraduates only in some of the problems (including the general domain problems), and they performed equally poorly as undergraduates in problems that required medical domain experience. There was no correlation between performance in the stock and flow problems and either duration of medical education or age. Thus we conclude that domain experience is not a strong indicator for overcoming the SF failure. Copyright © 2010 John Wiley & Sons, Ltd.

Syst. Dyn. Rev. (2010)

Introduction

The concept of accumulation is a broad concept applicable to problems in many different domains at the social, organizational and individual levels (Cronin *et al.*, 2009). All accumulation problems resemble a stock, additively increasing with inflows and decreasing with outflows over time. Even college students with strong technical backgrounds demonstrate difficulties when judging accumulation in very simple stock and flow problems (Booth Sweeney and Sterman, 2000; Cronin and Gonzalez, 2007; Cronin *et al.*, 2009; Sterman and Booth Sweeney, 2002). In multiple experiments, students answered questions about the stock and flow levels in the *People in the Store* task (PinS; Sterman, 2002; Cronin and Gonzalez, 2007; Cronin *et al.*, 2009), consisting of a line graph of inflow and outflow over a period of time. Most participants (more than 90 percent) answered flow questions correctly (questions 1 and 2), but not the stock questions (less than 50 percent correct for questions 3 and 4). This phenomenon has been called the "stock and flow failure" (SF failure) because participants' performance did not improve even after several interventions, including increasing the participants' motivation, reducing the number of data points in the graph, and varying the problem presentation (e.g. bar graphs, text) (Cronin *et al.*, 2009).

In Cronin *et al.* (2009), corrective feedback through repeated trials was the only intervention that helped improve performance. In that study, participants were asked to stay

[a] Department of Social and Decision Sciences, Carnegie Mellon University, 208 Porter Hall, Pittsburgh, PA 15213, U.S.A.
[b] University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, U.S.A.
[†] Author order is alphabetical and all authors contributed equally to the research reported in this paper.
* Correspondence to: Cleotilde Gonzalez. E-mail: coty@cmu.edu
Received 28 December 2009; Accepted 29 April 2010

in a room until they had answered the questions correctly or until one hour had passed, whichever came first. In each attempt, participants were told whether their answers were right or wrong. Participants' performance on the stock questions improved from 20 to 83 percent correct from the first to the ninth attempt. Although the difference in success rates of the stocks and flow questions was not statistically significant (Cronin *et al.*, 2009), the improvement is an indication that experience might help in solving these accumulation problems.

Our general goal was to determine the effects of domain experience in overcoming the SF failure. In this study, we investigated the effect of medical education for overcoming the SF failure in the medical domain. In medicine there are a number of situations that require physicians to judge the accumulation of a system in order to make decisions and keep the system in balance over time. For example, maintenance of fluid balance in a patient who is unable to take food or fluid by mouth involves monitoring the accumulation of fluids in the body over time by balancing inflows and outflows (see Figure 1 for an example of an isomorph of the PinS task in a medical context). The study of *domain* experience is a common approach to understanding expertise (Chi, 2006; Ericsson *et al.*, 1993), and there is an implicit assumption in the literature that those individuals who have greater experience in the domain may use more powerful heuristics that are relevant in the domain and that novices are not aware of (Chi, 2006). Since accumulation problems are very common in patient care we investigate the performance of medical students on medical and general accumulation problems, and compare their performance to that of a population of undergraduate students with no medical experience. If domain experience is a source of error in the SF problem, then:

> H1: the performance of medical students would be better in problems in the medical domain when compared to undergraduates with no medical education.

Methods

Participants

180 students at a school of medicine in the North East (age: $M = 26$ years, SD = 3.0; 99 male and 81 female; duration of medical education: $M = 1.5$ years, SD = 1.15) and 180 students (hereafter called undergraduates) from another non-medical school in the same city (age: $M = 23$ years, SD = 4.6; 112 male and 68 female) participated in this study. The undergraduates indicated several major areas of study, including Business Administration ($N = 42$), Engineering ($N = 24$), Finance ($N = 17$) Computer Science ($N = 11$), Economics ($N = 11$), Information Systems ($N = 13$), Business ($N = 8$), Music ($N = 8$), Mathematics ($N = 7$), Biology ($N = 6$), History ($N = 4$), Creative Writing ($N = 3$), other humanities ($N = 8$), and others ($N = 18$).

Stocks and flows task

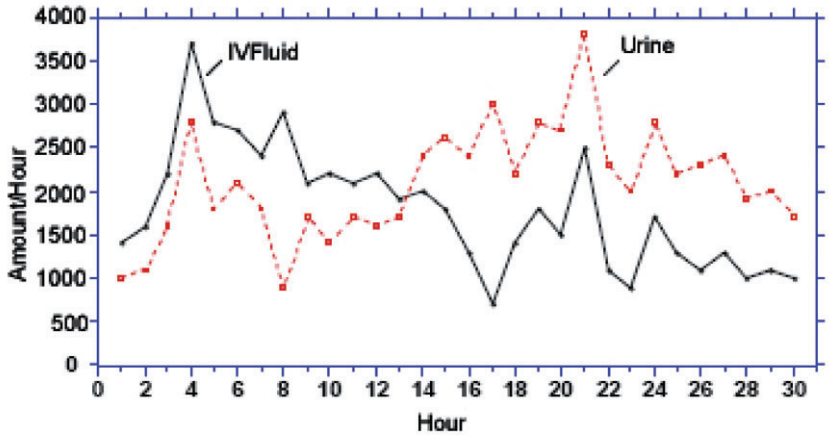
The task structure was adapted from the PinS task (e.g., Cronin *et al.*, 2009) into four medical and two generic versions of the same problem. Figure 1 shows the Fluids version (medical).

The medical versions of the PinS problem included *body fluids* resulting from IV fluids and urination (see Figure 1), *amount of bone tissue* resulting from osteoblast and

DOI: 10.1002/sdr

Fluids

The intake and output of fluids are often monitored, especially in a situation where a patient is unable to take fluids by mouth. The graph below shows the amount of intravenous fluid administered to a patient and the same patient's urine output over a 30 hour period. Assuming these are the only two factors affecting the amount of fluid in the body, please answer the following questions:



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

1. During which hour was the IV fluid highest?

   Hour …….. Can't be determined.

2. During which hour was the urine highest?

   Hour …….. Can't be determined.

3. During which hour were the body fluids highest (assumed all other factors are kept constant)?

   Hour …….. Can't be determined.

4. During which hour were the body fluids lowest (assumed all other factors are kept constant)?

   Hour …….. Can't be determined.

Fig. 1. An example of the Stocks and Flows Task for version 1 (fluids) as a medical version

osteoclast activity, *blood glucose level* resulting from glucagon and insulin production, and *temperature* resulting from attempts to stay warm and to cool down. For generic versions, we chose *weight* resulting from consumed and expended energy, and *PinS* resulting from people entering and leaving a store.

Versions differed in cover stories and labels, but had identical copies of the diagram. The correct responses for the flow questions are time points 4 and 21. To answer question 3 correctly, participants had to understand that the stock rises from time point 0

DOI: 10.1002/sdr

to 13 and falls thereafter, and thus the highest point of the stock is 13. To answer question 4 correctly, participants had to perceive that the area between the inflow and outflow is less from time point 0 to 13 than thereafter. Therefore, the stock is lowest at time point 30. Responses were coded as correct if they were between +1 and –1 time periods away from the correct answer.

Procedure

Participants gave informed consent, provided demographic information, and answered four questions for one task version. For the medical students, the study was conducted from 17 July 2008 to 18 December 2008, without compensation, using an email invitation and SurveyMonkey (http://www.surveymonkey.com). The email invitation reached all the medical students in the school, and this allowed us to obtain a high number of responders. For undergraduates, the study was conducted in September 2008 in the Student Center of the university, using paper and pencil and a candy bar as compensation. We excluded data from medical students who took more than 20 minutes to answer in the online survey. Each student included in this study (online and on paper) took an average of 5 minutes (SD = 8) to complete the task.

Results

Table 1 shows that, overall, medical students and undergraduates answered equally well to flow questions 1 and 2. In stock questions 3 and 4 all students answered poorly, but medical students responded more accurately than undergraduates overall. Although a general support of our hypothesis, this result varied considerably depending on the particular problem.

Chi-square analyses on proportion of correct responses for each of the two populations (see Table 1) revealed no significant differences between student populations for questions 1 and 2 (but see question 2 in version 6, PinS). For question 3, the correct responses for medical students ranged from 6.5 to 65.5 percent (M = 29.4 percent, SD = 25.1) and for undergraduates from 3.0 to 26.7 percent (M = 12.8 percent, SD = 12.9). Medical students performed significantly better than undergraduates on problems 1, 5, and 6. On question 4, the range of correct responses for medical students was 0–58.6 percent (M = 22.7 percent, SD = 23.5), and 0–26.7 percent (M = 13.9 percent, SD = 10.6) for undergraduates. Medical students performed significantly better than undergraduates on problem 6.

Some deeper analyses of the erroneous responses suggested evidence of the correlation heuristic as in Cronin *et al.* (2009). For question 3, 34 percent of the medical students (61 out of 180) responded with the time in which the net inflow was maximum (where the gap between the inflow and outflow is largest; t = 8), 16 percent (29 out of 180) responded with the time in which the outflow is maximum (t = 21); and 9 percent (17 out of 180) indicated that the answer "cannot be determined". For question 4, 34 percent of the medical students (62 out of 180) responded with the time in which the net outflow is maximum (t = 17); 16 percent (29 out of 180) responded with the time in which the inflow is maximum (t = 4; N = 55); and 14 percent (25 out of 180) responded

DOI: 10.1002/sdr

Table 1. Percentage of correct responses for the four questions for each of the six problems for both medical students (M) and undergraduates (U). The χ^2 test comparing both populations had 1 degree of freedom for all analyses. In the case of expected frequencies below 5, Fischer's exact test was performed

	N		Q1: Maximum inflow?				Q2: Maximum outflow?				Q3: Maximum stock?				Q4: Minimum stock?			
Problem	*N* (M)	*N* (U)	M	U	χ^2	*P*	M	U	χ^2	*p*	M	U	χ^2	*p*	M	U	χ^2	*p*
1 (fluids)	30	30	96.7	96.7	0.00	n.s.	96.7	96.7	0.00	n.s.	33.3	3.3	7.68	0.01	26.7	13.3	1.67	n.s.
2 (bone)	28	30	96.4	86.7	1.88	n.s.	96.4	93.3	0.29	n.s.	14.2	13.3	0.01	n.s.	7.1	16.7	1.28	n.s.
3 (glucose)	31	30	90.3	96.7	4.21	n.s.	90.3	93.3	0.19	n.s.	6.5	0.0	0.32	n.s.	0.0	0.0	—	—
4 (temp.)	30	30	100.0	100.0	—	n.s.	100.0	96.7	1.40	n.s.	6.7	3.3	0.36	n.s.	3.3	3.3	0.00	n.s.
5 (weight)	32	30	96.9	96.7	0.00	n.s.	96.9	96.7	0.00	n.s.	46.9	30.3	3.75	0.05	40.6	23.3	2.12	n.s.
6 (people)	29	30	96.6	96.7	0.00	n.s.	100.0	83.3	7.21	0.05	69.0	26.7	6.17	0.01	58.6	26.7	6.17	0.01
Total	180	180	96.1	95.6	0.00	n.s.	96.7	93.3	2.11	n.s.	29.4	12.8	14.45	0.001	22.8	13.9	4.75	0.05

DOI: 10.1002/sdr

with "cannot be determined". A similar but larger proportion of errors was found for undergraduates.

Point biserial correlation coefficients (Michael *et al.*, 1952) between correct performance for stock questions and duration of medical education and age revealed that correct performance was associated neither with duration of medical education, for question 3, $r_{pb} = 0.06$, $p = 0.22$, for question 4, $r_{pb} = 0.02$, $p = 0.39$, beyond their advantage compared to undergraduates, nor with age over all participants, for question 3, $r_{pb} = 0.03$, $p = 0.59$, for question 4, $r_{pb} = 0.04$, $p = 0.49$.

Discussion

The benefits of medical education in medical accumulation problems are supported by expertise research indicating that domain experience results in the use of relevant problem-solving strategies (Chi *et al.*, 1981; Chi, 2006). The overall advantage of domain experience when comparing medical and non-medical students is important, given that a considerable number of manipulations have been attempted with no success in improving responses to stock questions (Booth Sweeney and Sterman, 2000; Cronin and Gonzalez, 2007; Cronin *et al.*, 2009; Sterman and Booth Sweeney, 2002). However, despite the significant overall result, the advantage of domain experience comes from only a few of the problems, particularly the more generic problems (those in which medical students need to rely less on their medical domain experience). For example, it is surprising that medical students were no better than undergraduates in bone, glucose, and temperature problems, and even in the best of the cases, as in the fluid problem, in patient management one could not afford an optimistic 33.3 percent accuracy.

Error analysis indicates the use of an intuitive but erroneous heuristic termed the "correlation heuristic" (see Cronin *et al.*, 2009). Although to a less degree than undergraduates, medical students often ignore accumulation and make judgments based on differences in one point in time. As demonstrated by the correlation analyses, correct performance on stock questions does not relate to the duration of medical education or age.

Since accumulation problems are common in medical practice, the findings reported here have important implications for medical education, to reduce medical error, and to design systems that support physicians decisions. These results are a first hint at the role of medical education in solving stock and flow problems and they highlight the need to teach the concepts of accumulation in medical school.

In summary, our results indicate that domain experience is not a strong indicator for overcoming the SF failure. Several follow-up studies are needed in order to test the robustness of this result. First, given that the medical students used in this research had limited experience in the medical field (our sample showed a mean of 1.5 years of medical education), it is possible that medical students were not, after all, much different from undergraduates. It would be interesting to test participants who vary more widely on medical experience. Second, given the different methods used for the two main groups of concern (medical students were tested online while undergraduates were tested on paper), it would be important to replicate these results in a study where exactly the same methods are used for the different groups. Third, the replication of

DOI: 10.1002/sdr

this study in other domains of knowledge would also help in investigating the robustness of these findings.

Acknowledgements

This research was partially supported by the National Science Foundation (Human and Social Dynamics: Decision, Risk, and Uncertainty, Award No. 0624228) award to Cleotilde Gonzalez. We want to thank members of the Dynamic Decision Making Laboratory, Varun Dutt for the implementation of the problems on surveymonkey.com, and Hau-yu Wong for her editorial review of this paper. We also thank Eliza Beth Littleton from the School of Medicine at the University of Pittsburgh for her helpful comments in previous versions of this manuscript. An earlier version of this work was reported at the System Dynamics Meeting by Gonzalez, Brunstein, and Kanter (2009).

Biographies

Angela Brunstein, Ph.D., is a Post-Doctoral Teaching Fellow at the Qatar campus of Carnegie Mellon University. Before her current position she was a post-doctoral fellow at the Dynamic Decision Making Laboratory. She has conducted extensive research in complex problem solving and learning. Her areas of expertise include e-learning. Dr. Brunstein teaches introduction to psychology and cognitive psychology.

Cleotilde (Coty) González, Ph.D., is an Associate Research Professor at the Department of Social and Decision Sciences at Carnegie Mellon University in Pittsburgh, PA. She is the founding director of the Dynamic Decision Making Laboratory where researchers conduct behavioral studies on dynamic decision making using Decision Making Games, and create technologies and cognitive computational models to support decision making and training. Her research work focuses on the study of human decision making in dynamic and complex environments.

Steven L. Kanter, MD, serves as the Vice Dean of the University of Pittsburgh School of Medicine (UPSOM). Dr. Kanter draws from a diverse background of experience that includes clinical medicine, medical informatics, medical education, scholarly publishing, and medical school administration. He has been a driving force in curricular renewal at UPSOM, has played a key role in reformulating guidelines for promotion of faculty, and has established a system of "promotion pathways" at UPSOM which provides an explicit framework for career development.

References

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

Chi MTH. 2006. Two approaches to the study of experts' characteristics. In *The Cambridge Handbook of Expertise and Expert Performance*, Ericsson KA, Charness N, Feltovich PJ, Hoffman RR (eds). Cambridge University Press: New York; 21–30.

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

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

DOI: 10.1002/sdr

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**(1): 116–130.

Ericsson KA, Krampe RT, Tesch-Roemer C. 1993. The role of deliberate practice in acquisition of expert performance. *Psychological Review* **100**: 363–406.

Gonzalez C, Brunstein A, Kanter SL. 2009. On the role of medical experience for overcoming the stocks and flows failure. In *Proceedings of the 27th International Conference of the System Dynamics Society, Albuquerque, NM*.

Michael WB, Perry NC, Guilford JP. 1952. The estimation of a point biserial coefficient from a phi coefficient. *British Journal of Statistical Psychology* **5**: 139–150.

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

Sterman JD, Booth Sweeney L. 2002. Cloudy skies: assessing public understanding of global warming. *System Dynamics Review* **18**(2): 207–240.

DOI: 10.1002/sdr