Mathematical knowledge is related to understanding stocks and flows: results from two nations

Liang Qi[a,b] and Cleotilde Gonzalez[a*]

Abstract

Stocks and flows (SF) are essential to every-day judgments: debt changes with rates of incomes and expenses, and body weight with calories consumed and energy expended. Research suggests that individuals with strong mathematical skills have a poor understanding of SF. However, past research used homogeneous participant samples and the relationship between mathematical knowledge and performance in SF tasks was not tested. In two studies involving different populations from China and the U.S.A, we find that individuals with better general mathematical knowledge tend to be more accurate in SF tasks. We find that most participants who make mistakes follow an erroneous correlation heuristic; however, we also find that the use of this heuristic is not related to mathematical knowledge. Our results open the door to new research questions, including the type of mathematical knowledge needed and the relationship of mathematical knowledge and cognitive processes that people need for solving SF tasks. Copyright © 2015 System Dynamics Society

Syst. Dyn. Rev. (2015)

Introduction

Stocks and flows (SF) are essential building blocks of every dynamic system, from the accumulation of water in a bathtub to the accumulation of greenhouse gases in the atmosphere. Effective decision making in dynamic systems requires an understanding of the process of accumulation through inflows and outflows over time. For example, inventory managers should understand warehouse stock levels, the upstream production rate and the product usage rate to know when to replenish; physicians must understand changes in a patient's body fluids resulting from intravenous therapy and urination to provide a better fluid therapy plan; and people have to manage their daily caloric intake and energy expenditure to maintain a healthy weight. However, research has documented a robust problem in human reasoning: our inability to correctly judge the process by which inflows and outflows accumulate over time (*SF failure*; Cronin *et al*., 2009). Many studies conclude that highly educated individuals in well-known technical universities perform poorly even

[a] Dynamic Decision Making Laboratory, Department of Social and Decision Sciences, Carnegie Mellon University, 4609 Winthrop Street, Pittsburgh, PA 15213, U.S.A

[b] Department of Health Service, Second Military Medical University, Shanghai, China

* Correspondence to: Cleotilde Gonzalez, Dynamic Decision Making Laboratory, Department of Social and Decision Sciences, Carnegie Mellon University, 4609 Winthrop Street, Pittsburgh, PA 15213, U.S.A. E-mail: coty@cmu.edu

Accepted by Luis Luna-Reyes, *Received 29 January 2015*; *Revised 22 September 2015*; *Accepted 30 September 2015*

System Dynamics Review
System Dynamics Review (2015)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/sdr.1539

in extremely simplified SF tasks (Sweeney and Sterman, 2000; Sterman, 2002; Sterman and Sweeney, 2002; Cronin *et al.*, 2009). Research demonstrates that the SF failure is not an artifact of the task, that it is not easy to correct, and that it reflects serious misunderstandings of basic principles of accumulation (Cronin *et al.*, 2009). Instead, in solving SF problems a majority of participants use a *correlation heuristic*, a form of erroneous linear reasoning where participants expect the behavior of a stock to be similar to that of the flow (Cronin *et al.*, 2009). However, little progress has been made on determining the factors that lead to the SF failure and correlational reasoning.

Recently, some progress toward understanding some of the cognitive underpinnings indicates that the SF failure appears when people focus on specific system elements (local processing), rather than on the system structure and gestalt (global processing) (Fischer and Gonzalez, 2015): when individuals possess such ability for global processing, they seem more successful at SF tasks than those with local processing ability. Similarly, several studies have found that analytical thinking style relates to better performance in SF problems compared to intuitive thinking style (Lakeh and Ghaffarzadegan, 2015; Weinhardt *et al.*, 2015) and this effect has also been linked recently to a broader effect in multi-echelon supply chains (Narayanan and Moritz, 2015). Other researchers find that the individual's familiarity with the context (Newell *et al.*, 2015) and the similarity of the surface and deep features may make it more or less difficult to process SF tasks (Gonzalez and Wong, 2012).

However, in a large majority of studies on SF failure, researchers have tested relatively homogeneous populations (e.g. students in American universities) with a strong background in mathematics (exceptions are Abdel-Hamid *et al.*, 2014; Lakeh and Ghaffarzadegan, 2015; Fischer and Gonzalez, 2015), and the link between mathematical skills and those cognitive factors has not been made. Researchers often suggest that mathematical education may be related to performance in simple SF problems by pointing out that "highly educated subjects with extensive training in mathematics and science" (Sweeney and Sterman, 2000, p. 278) have a poor understanding of some of the most basic concepts of dynamic systems, but the link between mathematics and understanding of stocks and flows has not been shown. Although it may seem straightforward, it is difficult to determine how mathematical education may relate to SF task performance and to the type of correlational behavior that people exhibit in these tasks more generally. Recently it was found that mathematical background was not a significant predictor for answering correctly to the SF tasks, but no explanation for this result was provided. For example, the relationships between mathematical knowledge and analytical mode of thinking were not tested (Lakeh and Ghaffarzadegan, 2015). In addition, the researchers used a categorical, self-reported value of "mathematical expertise", instead of objective measures of mathematical knowledge and skills.

In the current research, we formally address the relationships between mathematical knowledge and SF failure and correlational behavior in two large-scale studies in two nations (China and the U.S.A.), which are known to differ in their

DOI: 10.1002/sdr

approach towards mathematical education. Chinese students are well regarded for their mathematical achievements and abilities, while U.S. counterparts are often said to lag behind (Stevenson *et al.*, 1986; Beaton, 1996; Mullis *et al.*, 1997; Grow-Maienza *et al.*, 2001; Leung, 2006). Also, a recent study suggests comparatively lower performance in an SF task for U.S. participants relative to Chinese participants (Abdel-Hamid *et al.*, 2014). Generally speaking, we expect to demonstrate a positive relationship between basic mathematical knowledge and SF performance when diverse participant samples are tested. Given a common reliance of a correlation heuristic, we also expect that mathematical knowledge will be related to the use of this heuristic in solving the problems.

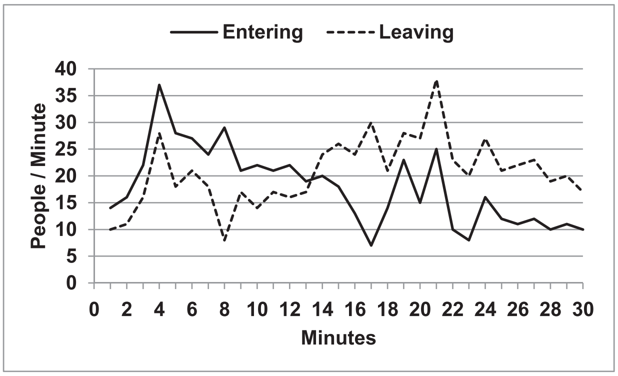
SF failure in the department store task

Investigations of the SF failure emerge from a need to understand human behavior and decision making in dynamically complex environments (Sweeney and Sterman, 2000). Human decision making in complex dynamic environments is often poor, and learning is slow and weak even when learning conditions are facilitated (e.g. Sterman, 1989a, 1989b; Paich and Sterman, 1993; Diehl and Sterman, 1995; Gonzalez, 2004; Gonzalez *et al.*, 2005). In an attempt to identify the issues that lead to SF failure, researchers reduced the complexity of dynamic systems dramatically; yet, despite that, these problems with the SF failure remained (Sweeney and Sterman, 2000; Sterman, 2002; Cronin and Gonzalez, 2007).

A large number of simple SF tasks were developed to assess basic systems thinking concepts: stocks and flows, time delays and negative feedback (Sweeney and Sterman, 2000). The “department store” task (DS task) (Sterman, 2002) presents participants with a graph showing the number of people entering (inflow) and leaving (outflow) a department store each minute over a 30-minute interval (Figure 1). To assess participants’ understanding of stocks and flows, they are asked four questions. Q1 and Q2 test whether participants can read the graph and correctly distinguish between inflow and outflow; Q3 and Q4 test whether participants can infer the behavior of the stock from the behavior of the flows. Although a majority of people typically answer Q1 and Q2 correctly (more than 90%), a minority can answer Q3 and Q4 correctly (less than 50%) (Cronin and Gonzalez, 2007; Cronin *et al.*, 2009). This difficulty in understanding stock–flow dynamics seems to stem from the erroneous tendency to perceive a stock’s behavior as directly related to that of its flows (the *correlation heuristic*; Cronin *et al.*, 2009).

Mathematically speaking, a stock at any time is defined as the integral of its net inflow over time plus the quantity in stock at the initial time (Cronin *et al.*, 2009), but an answer to Q3 and Q4 could be obtained by manually tallying the number of people in the store, where the number of people at time t is the accumulation of previous time periods plus the inflow minus the outflow at that point of time ($S_t = S_{t-1} + I_t - O_t$). However, these calculations are not necessary if one understands the basic principles of accumulation: that the number

DOI: 10.1002/sdr



Based on the information provided in the graph above, please answer the following questions with a number from 1 to 30. Check the box if you feel that the answer cannot be determined with the information provided.

Q1: At which minute did the most people enter the store?

Minute_________ ☐ Can't be determined.

Q2: At which minute did the most people leave the store?

Minute_________ ☐ Can't be determined.

Q3: After which minute were the most people in the store?

Minute_________ ☐ Can't be determined.

Q4: After which minute were the fewest people in the store?

Minute_________ ☐ Can't be determined.

Fig. 1. Department store (DS) task as used in Sterman (2002) and in some of the experiments in Cronin *et al.* (2009)

of people in the store rises when the flow of people entering is greater than the flow of people leaving (and vice versa), and then note that the number entering is greater than the number exiting through time 13 and less thereafter. We do not know whether this is an intuition emerging from basic mathematical training. Our initial goal addressed here is to establish a formal link between general and basic mathematical knowledge and performance in this type of SF task.

Study 1: Chinese high school students

Chinese students are well regarded for their mathematical achievements and abilities (Stevenson *et al.*, 1986; Beaton, 1996; Mullis *et al.*, 1997; Grow-Maienza *et al.*, 2001; Leung, 2006), and they are often among the top performers in mathematics by international standards (Fan and Zhu, 2004). Given that we had access to Chinese high school students after they had been tested for their mathematical skills, we were able to test Chinese participants from groups of different mathematical skills in the SF tasks.

DOI: 10.1002/sdr

Methods

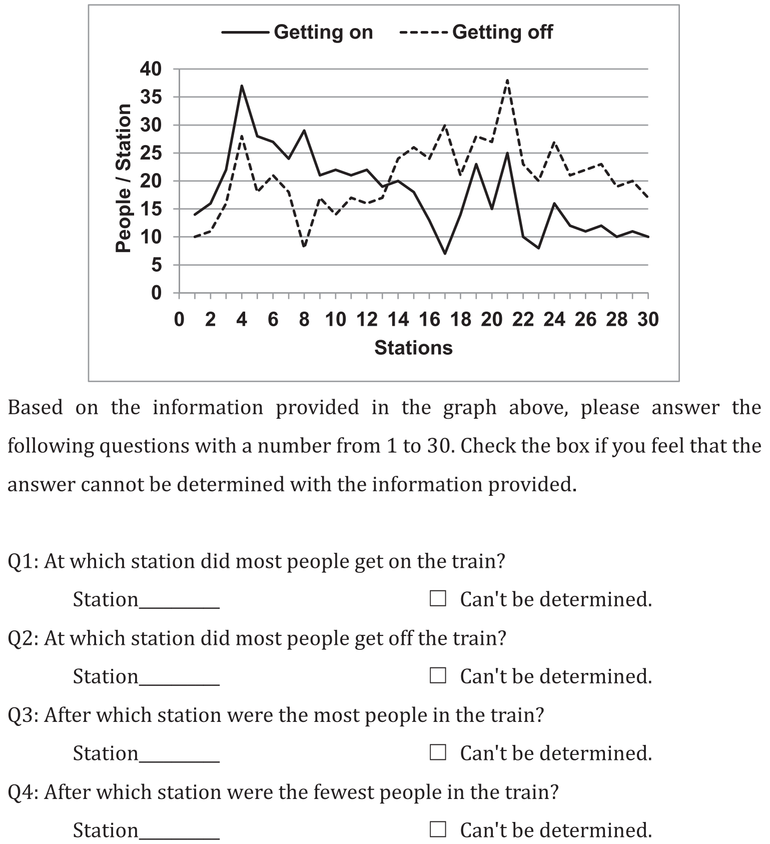
Participants

We recruited 224 students from a high school in mainland China (age: $M = 16.98$ years, $SD = 0.44$; 95 men and 129 women). The high school chosen contains a typical population of students taking National Higher Education Entrance Examination (commonly known as “Gaokao” in China) every year. The high school administration was contacted about our experiment, and they volunteered to give supervised access to students in the classrooms. No class credit was given for participation.

Materials and procedures

We used the DS task (Figure 1) and an isomorphic representation of this task, the Metro/Train task (MT task in Figure 2). The MT task was used to provide some diversity in the problems given and to informally explore the claims

Fig. 2. The metro train (MT) task: an isomorph of the DS task



Based on the information provided in the graph above, please answer the following questions with a number from 1 to 30. Check the box if you feel that the answer cannot be determined with the information provided.

Q1: At which station did most people get on the train?

Station_________ □ Can't be determined.

Q2: At which station did most people get off the train?

Station_________ □ Can't be determined.

Q3: After which station were the most people in the train?

Station_________ □ Can't be determined.

Q4: After which station were the fewest people in the train?

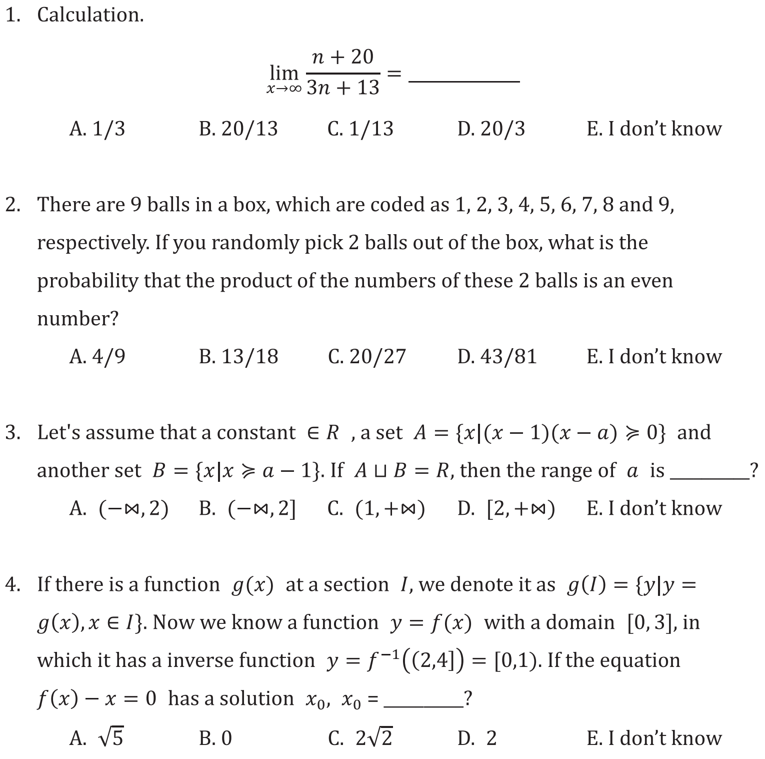
Station_________ □ Can't be determined.

DOI: 10.1002/sdr

recently made regarding context familiarity that may be beneficial for understanding SF (Newell *et al.*, 2015). We expected the MT task to be a context with which the Chinese high school students would be more familiar. They take the metro/train every day in the city, and it may be easier for them to make a connection with the concepts of accumulation (e.g. a train “holds” people, and people get in and out of the train). The DS and MT tasks were isomorphs.

All participants answered a four-question mathematics test (Figure 3), which was taken from the Gaokao test in Shanghai (2013 Shanghai Mathematics Test, 2014). The score in the mathematical test ranged from 0 to 4: the number of correct responses.

First, all participants answered a demographic questionnaire, including age, gender, nationality and educational background. Next, participants answered the mathematics test (Figure 3). Then participants were randomly assigned to do one SF task: either the DS task or the MT task. They were asked to read the problem and to answer the questions. They were not given any time limit and were free to take as long as they needed. The whole task was finished in 15 minutes, and all materials were collected afterwards.



1. Calculation.

$$\lim_{x\to\infty}\frac{n+20}{3n+13} = \_\_\_\_\_\_\_\_$$

A. 1/3 B. 20/13 C. 1/13 D. 20/3 E. I don’t know

2. There are 9 balls in a box, which are coded as 1, 2, 3, 4, 5, 6, 7, 8 and 9, respectively. If you randomly pick 2 balls out of the box, what is the probability that the product of the numbers of these 2 balls is an even number?

A. 4/9 B. 13/18 C. 20/27 D. 43/81 E. I don’t know

3. Let's assume that a constant $\in R$, a set $A = \{x|(x-1)(x-a) \geqslant 0\}$ and another set $B = \{x|x \geqslant a-1\}$. If $A \sqcup B = R$, then the range of a is ________?

A. $(-\bowtie, 2)$ B. $(-\bowtie, 2]$ C. $(1, +\bowtie)$ D. $[2, +\bowtie)$ E. I don’t know

4. If there is a function $g(x)$ at a section I, we denote it as $g(I) = \{y|y = g(x), x \in I\}$. Now we know a function $y = f(x)$ with a domain $[0, 3]$, in which it has a inverse function $y = f^{-1}\big((2,4]\big) = [0,1)$. If the equation $f(x) - x = 0$ has a solution x_0, x_0 = ________?

A. $\sqrt{5}$ B. 0 C. $2\sqrt{2}$ D. 2 E. I don’t know

Fig. 3. Mathematics test questions (correct answers are 1-A, 2-B, 3-B, 4-D)

DOI: 10.1002/sdr

Responses to each question were coded as correct, if: Q1 = 3, 4 or 5; Q2 = 20, 21 or 22; Q3 = 12, 13 or 14; Q4 = 29, 30. That is, answers were considered correct if they were within ±1 of the correct response, as done in past studies (Cronin and Gonzalez, 2007; Cronin *et al.*, 2009; Gonzalez and Wong 2012).

Results

Table 1 shows the distribution of points for Chinese respondents in the mathematical test. The table shows the number of participants that answered zero, one, two, three and four questions correctly. About 16 percent (16.5 percent) of Chinese students answered all questions wrong in the mathematics test, and 17.9 percent answered three questions correctly. The majority of participants answered one or two questions correctly.

Table 2 shows overall performance rates of SF tasks. Over half (55.4 percent) of Chinese participants answered Q3 correctly, and 56.3 percent answered Q4 correctly. The scores of the Chinese high school students were better than those of highly educated samples reported in Sterman (2002) (42 percent of Q3 and 30 percent of Q4); higher than those reported in Cronin *et al.* (2009) (44 percent of Q3 and 31 percent of Q4) from students enrolled in a graduate course in systems thinking; and higher than those reported in Cronin and Gonzalez (2007) (41 percent of Q3 and 33 percent of Q4) from undergraduate students in a private North American university.

Table 1. Distribution of Chinese respondents who answered none, one, two, three or all four of the mathematical test questions correctly

	Score				
	0	1	2	3	4
China ($N = 224$)	37 16.5%	75 33.5%	72 32.1%	40 17.9%	0 00.0%

Table 2. Overall performance rates in the SF tasks by Chinese respondents. Performance rates between MT and DS respondents are compared with χ^2 tests (d.f. = 1); significant p-values at *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$ are in bold type

	SF task question			
SF task	Q1	Q2	Q3	Q4
MT ($N = 112$)	107 95.5%	104 92.9%	72 64.3%	76 67.9%
DS ($N = 112$)	104 92.9%	104 92.7%	52 46.4%	50 44.6%
χ^2	0.327	0	**6.521***	**11.338*****
All ($N = 224$)	211 94.2%	208 93.3%	124 55.4%	126 56.3%

DOI: 10.1002/sdr

Table 2 also summarizes the performance rates in the two isomorphic representations of the SF task. Interestingly, the MT representation of the SF task resulted in more accurate responses than the DS task in both Q3 ($p < 0.05$) and Q4 ($p < 0.01$).

SF performance and mathematical scores

We built logistic regression models to estimate the probability of answering Q3 and Q4 correctly using three factors: age, gender and mathematical score. Table 3 shows that the log of the odds of Q3 and Q4 being correct is positively related to the mathematical score. In other words, the higher the mathematical score, the more likely it is that a Chinese participant would give the correct answer for each of these questions. Age and gender were not significant predictors of the correctness of Q3 or Q4.

Correlational thinking

Analogously to previous research (Cronin and Gonzalez, 2007; Cronin *et al.*, 2009), answers were coded as matching a correlation heuristic when participants answered in minute 4 (max. inflow) or minute 8 (max. net inflow) in Q3, and in minute 21 (max. outflow) or minute 17 (max. net outflow) in Q4. Table 4 shows the results of correlation heuristic use and other errors in the Chinese group. Among all the participants, the most common error is the correlation heuristic. In the Chinese sample, 71.0 percent of students who did not answer correctly chose to follow a correlation heuristic for Q3; and 71.4 percent of students for Q4.

We also built logistic regression models to estimate the probability of participants' following a correlation heuristic when they could not answer Q3 and Q4 correctly using age, gender, and mathematical score as predictors. Table 5 shows that the log of the odds of following a correlation heuristic for erroneous answers to Q3 or Q4 is positively related to gender, but not to

Table 3. Logistic regression models for correctness of Q3 and Q4 predicted by age, gender (1 = male; 2 = female) and mathematical score. Significant *p*-values at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ are in bold type

	Estimate	SE	Odds ratio
Q3 (maximum accumulation)			
Mathematical score	**1.979****	0.262	7.236
Age	0.031	0.426	1.032
Gender	0.094	0.360	1.098
Constant	−3.402	7.330	0.033
Q4 (minimum accumulation)			
Mathematical score	**0.998****	0.328	2.713
Age	1.160	0.654	3.191
Gender	−0.274	0.556	0.760
Constant	**−23.844***	11.367	0.000

DOI: 10.1002/sdr

Table 4. Distributions of errors for Chinese respondents. Bold values highlight evidence of correlational thinking

Error type	Q3 error distribution	Q4 error distribution
Max. inflow ($t = 4$)	**4** (**4.0**%)	0 (0.0%)
Max. net inflow ($t = 8$)	**67** (**67.0**%)	1 (1.0%)
Max. outflow ($t = 21$)	2 (2.0%)	**3** (**3.1**%)
Max. net outflow ($t = 17$)	15 (15.0%)	**67** (**68.4**%)
Other	9 (9.0%)	22 (22.5%)
Don't know	3 (3.0%)	5 (5.1%)
Overall	100 (100.0%)	98 (100.0%)
Correlation heuristic	**71** (**71.0**%)	**70** (**71.4**%)

Table 5. Logistic regression model for following a correlation heuristic in Q3 and Q4, predicted by age, gender (1 = male; 2 = female) and mathematical score. Significant p-values at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ are in bold type

	Estimate	SE	Odds ratio
Q3 (*maximum accumulation*)			
Mathematical score	−0.062	0.157	0.940
Age	−0.384	1.336	0.681
Gender	**0.907****	3.315	2.477
Constant	4.377	5.725	79.630
Q4 (*minimum accumulation*)			
Mathematical score	−0.163	0.159	0.850
Age	0.018	0.338	1.018
Gender	**1.014****	0.321	2.756
Constant	−2.497	5.787	0.082

mathematical score or age. Since we coded men as 1 and women as 2 in the dataset, the result suggests that female participants are more likely to rely on the correlation heuristic when they answer incorrectly Q3 and Q4 than do male participants.

Summary of results

Overall, the SD task performance of the Chinese high school students was better than that of highly educated samples in previous studies, and the MT representation of the SF task resulted in more accurate responses than the DS task. Furthermore, higher scores in the SF task and answering questions 3 and 4 correctly were related to higher scores in the mathematical test. Interestingly, analyses of mistakes made in the accumulation questions indicate that a large majority of errors were made by following a correlation heuristic. These errors are not predicted by their mathematical score. Gender was the only statistically significant variable, where female participants were more likely to use a correlation heuristic.

Study 2: diverse U.S. Sample

With a few exceptions (e.g. Abdel-Hamid *et al*., 2014; Fischer and Gonzalez, 2015; Lakeh and Ghaffarzadegan, 2015), most studies of the SF failure have relied on homogeneous samples of highly educated students in the U.S.A. Our Study 1 introduces a different population of high school students from China. However, this group of participants is still homogeneous in age and relatively similar in mathematical knowledge given their similar educational background. Study 1 suggested a positive relationship between mathematical knowledge and accuracy in the SF questions and the prevalence of the correlation heuristic, prominently committed by female participants rather than male participants in a sample of Chinese high school students. But a question still remains about the effect of mathematical knowledge in a more diverse sample, and arguably, also a more representative sample of the U.S. population. This is a question addressed in this study.

Methods

Participants

Two hundred and forty-three U.S. participants were recruited via Amazon Mechanical Turk online (age: $M = 34.65$ years, $\text{SD} = 12.13$; 158 men and 85 women). They were restricted to only U.S. IP addresses. About 53.7 percent had a 4-year college course as highest level of education and 13 percent (13.6 percent) had high school.

Materials and procedures

The same mathematics test and SF tasks as in Study 1 were used in this study. The two versions of the tasks and the mathematical test were designed in Qualtrics. MTurk workers got $0.50 for completing the study. They were given extra incentive to do well in the task: if they correctly answered all of the mathematical questions, they would get a $0.50 bonus, and if they correctly answered all of the SF task questions, they would also get an additional $0.50 bonus. Participants completed the same demographic questionnaire, mathematical test, and one of the DS or MT tasks. They were not given any time limit and were free to take as long as needed. U.S. participants spent an average of 3.9 minutes (range: 7.96 seconds to 44.50 minutes) on the mathematical test, and 1.9 minutes (range: 1.92 seconds to 16.67 minutes) for the SF task.

Results

Table 6 shows the distribution of points of U.S. respondents in the mathematical test. About half of the American participants (46.1 percent)

DOI: 10.1002/sdr

Table 6. Scores of each mathematical test question for U.S. Respondents

	Score				
	0	1	2	3	4
U.S. ($N = 243$)	112 46.1%	92 37.8%	33 13.6%	4 1.7%	2 0.8%

answered all of the questions in the mathematical test wrong, and 37.9 percent answered one question correctly. Thus the majority of participants answered none or one mathematical question correctly, despite the extra incentive.

Table 7 shows overall performance rates of SF tasks. Only 16.5 percent of U.S. participants answered Q3 correctly, and 18.9 percent answered Q4 correctly. The scores of the MTurk workers were lower than any of those reported in past research: Sterman (2002) (42 percent of Q3 and 30 percent of Q4), Cronin *et al.* (2009) (44 percent of Q3 and 31 percent of Q4), Cronin and Gonzalez (2007) (41 percent of Q3 and 33 percent of Q4); but these are similar to the rates found in other studies conducted on MTurk recently (Fischer and Gonzalez, 2015). These differences might have several reasons. Previous experiments were mostly conducted in classrooms and laboratories, which provided quiet environments without distractions. With the MTurk participants, as with most Internet-based methods, experimenters can exert only minimal control over participants' environments compared to lab studies (Buhrmester *et al.*, 2011). Also, previous studies only recruited highly educated participants who were from colleges and graduate schools, while MTurk participants are more socio-economically and ethnically diverse (Casler *et al.*, 2013). Although 53.7 percent received 4-year college education and above, the time and place in which they received their degree are less homogeneous than in past research populations.

Table 7. Overall performance rates in the SF tasks by U.S. Respondents. Performance rates between MT and DS respondents were compared with χ^2 tests (d.f. = 1). none of the tests was significant ($p > 0.05$). when χ^2 appro-ximation was inadequate due to small expected frequency (Q1), the *p*-value of Fisher's exact test was used and it was not significant ($p > 0.05$)

	SF task question			
Nation	Q1	Q2	Q3	Q4
MT ($N = 118$)	114 96.6%	111 94.1%	14 11.9%	20 16.9%
DS ($N = 125$)	117 93.6%	113 90.4%	26 20.8%	26 20.8%
χ^2	—	0.681	2.905	0.362
All ($N = 243$)	231 95.1%	224 92.2%	40 16.5%	46 18.9%

DOI: 10.1002/sdr

Table 7 also summarizes the performance rates in the two isomorphic representations of the SF task. The two representations did not result in different success rates for all four questions.

SF performance and mathematical scores

Logistic regression models were built to estimate the probability of answering Q3 and Q4 correctly, using age, gender, and mathematical score as predictors. Table 8 shows that the log of the odds of answering Q3 and Q4 correctly is positively related to mathematical score and negatively related to age and gender. These results indicate that the higher the mathematical score, the more likely it is that a participant would give the right answer for each of the two questions. Male participants were more likely to perform better than female participants, and younger participants responded more accurately than older ones.

Correlational thinking

Table 9 shows the results regarding correlation heuristic use and other errors in the American group. Among all of the participants, most of the incorrect responses are due to a common error of following the correlation heuristic. In the American population, 59.1 percent of participants, who did not answer correctly, chose to follow a correlation heuristic for Q3; and 44.7 percent of participants for Q4. These proportions are lower than among the Chinese sample, mostly due to an increase of “Other” and “Don’t know” responses within the U.S. sample.

Again, logistic regression models were used to estimate the probability of MTurk workers’ following a correlation heuristic when they could not answer Q3 and Q4 correctly using age, gender, and mathematical score as predictors. In these models reported in Table 10, there is no effect of gender, age, or mathematical score. Thus, when answering erroneously, the use of the correlation heuristic cannot be predicted by any of these factors.

Table 8. Logistic regression model for Q3 and Q4 correctness of MTurk participants predicted by age, gender (1 = male, 2 = female) and mathematical score. Significant p-values at $*p < 0.05$, $**p < 0.01$, $***p < 0.001$ are in bold type

	Estimate	SE	Odds ratio
Q3 (*maximum accumulation*)			
Mathematical score	**0.471*****	0.114	1.601
Age	−**0.051*****	0.012	0.950
Gender	−**0.045***	0.213	0.638
Constant	−0.712	0.514	2.037
Q4 (*minimum accumulation*)			
Mathematical score	**0.584*****	0.119	1.794
Age	−**0.061*****	0.013	0.941
Gender	−**0.749*****	0.221	0.473
Constant	**1.318***	0.536	3.735

DOI: 10.1002/sdr

Table 9. Distributions of errors for U.S. Respondents. Bold values highlight evidence of correlational thinking

Error type	Q3 error distribution	Q4 error distribution
Max inflow ($t = 4$)	**29**(**14.3**%)	0(0.0%)
Max net inflow ($t = 8$)	**91**(**44.8**%)	8(4.1%)
Max outflow ($t = 21$)	10(4.9%)	**8**(**4.1**%)
Max net outflow ($t = 17$)	13(6.4%)	**80**(**40.6**%)
Other	34(16.7%)	65(33.0%)
Don't know	26(12.8%)	36(18.3%)
Overall	203(100.0%)	197(100.0%)
Correlation heuristic	**120**(**59.1**%)	**88**(**44.7**%)

Table 10. Logistic regression model for following a correlation heuristic in Q3 and Q4, predicted by age, gender (1 = male; 2 = female) and mathematical score for U.S. Participants. None of the values was significant ($p > 0.05$)

	Estimate	SE	Odds ratio
Q3 (*Maximum accumulation*)			
Mathematical score	−0.073	0.161	0.929
Age	−0.001	0.011	0.999
Gender	−0.440	0.278	0.644
Constant	0.661	0.530	1.937
Q4 (*minimum accumulation*)			
Mathematical score	−0.245	0.173	0.783
Age	−0.013	0.012	0.988
Gender	−0.113	0.290	0.894
Constant	0.192	0.554	1.211

Summary of results

Overall, the performance scores of the MTurk workers were lower than those of Chinese students and those of students in past studies with highly educated U.S. participants. Performance was similar in both versions of the SF task: the MT and the DS tasks. Higher accuracy in the accumulation questions was predicted by higher scores in the mathematical test, and lower accuracy was predicted by female respondents. Also, younger participants were more accurate compared to older participants. Most errors committed were the correlation heuristic, but age, gender and mathematical score were not related to the use of the correlation heuristic.

Discussion and conclusions

Our results contribute to improved understanding of the factors that explain the SF failure and the errors committed when solving SF tasks. Generally speaking, and in agreement with the growing literature, our results show that across the two studies participants perform poorly in SF tasks. However, our results also show interesting variability in accuracy. Chinese high school

DOI: 10.1002/sdr

students were more accurate at SF questions than students in past studies (e.g. Cronin *et al.*, 2009), who were similarly homogeneous but from highly technical universities in the U.S.A. Also, in agreement with some recent studies (Fischer and Gonzalez, 2015; Lakeh and Ghaffarzadegan, 2015), the largely heterogeneous population of MTurk US workers did extremely poorly on the SF questions. This is worrisome, because arguably the MTurk sample may be more representative of the general American population compared to students from elite universities. Furthermore, in agreement with other studies (Sterman, 2010; Lakeh and Ghaffarzadegan, 2015), we find a correlation between age, gender and performance in the SF tasks, with males performing more accurately than females, and younger participants responding more accurately than older ones. We also find that the MT task, an isomorph of the DS task, results in better performance for the Chinese participants, who are expected to be familiar with this context, given that most people in China are familiar with trains, the country's primary mode of transport. The MT task results in no different performance as the DS task for the U.S. sample. The effects of the familiarity of the context on performance in these tasks have been investigated widely but the results seem inconclusive (Cronin and Gonzalez, 2007; Brunstein *et al.*, 2010; Newell *et al.*, 2015). Our results provide some partial support about the positive effects of familiarity on the SF failure.

More importantly to the point of this research, this variability in the participant samples is essential in testing a relationship between mathematical knowledge (and other demographic characteristics) and SF task performance. Across the two studies we find that people that score higher in a general mathematical test are able to respond more accurately to accumulation tasks than those with lower scores. Given the homogeneity of the samples used in past studies, the assumption was that students with high levels of mathematical education were unable to do well in these tasks (Sweeney and Sterman, 2000; Cronin *et al.*, 2009). However, we now show that individuals vary in their levels of performance and that their mathematical knowledge matters. This result may seem to contradict a recent finding (Lakeh and Ghaffarzadegan, 2015) from a similar MTurk diverse population, which found no significant correlation between mathematical skills and performance in SF tasks; but their study involved self-reported levels of "mathematical expertise", a less objective measure than asking participants to solve mathematical problems.

Yet the reasons for which mathematical knowledge is related to SF performance are not known. Cronin *et al.* (2009) suggested that early mathematical education may inadvertently reinforce the SF failure and recent studies (e.g., Dutt and Gonzalez, 2012) found a significant relationship between education in science, technology, engineering and mathematics (STEM) and the SF failure in judgments of CO_2 accumulation. However, these studies provide limited answers to a more interesting question: how and why does mathematical

DOI: 10.1002/sdr

knowledge and education influence our understandings of SF and dynamic systems more generally?

Research in mathematical education suggests that students receive extensive practice in proportional reasoning that may encourage linear thinking and reinforce the impression that relations between variables are proportional (Brink and Streefland, 1979; Van Dooren *et al.*, 2005; De Bock *et al.*, 2007), failing to communicate and demonstrate the principles of accumulation (Behr *et al.*, 1992; Ben-Zeev and Star, 2001; Van Dooren *et al.*, 2005). However, our research finds no significant relationships between mathematical knowledge and the use of the correlation heuristic.

Recent studies have suggested that participants' cognitive mode of thinking, whether analytical or intuitive, might influence performance in SF tasks and supply chain management decisions (Lakeh and Ghaffarzadegan, 2015; Narayanan and Moritz, 2015). However, these studies do not test the connection of these cognitive factors and mathematical knowledge. A simple intervention to encourage analytical thinking was successful in improving performance in SF tasks but there was no effect of such interventions in the use of correlational thinking (Lakeh and Ghaffarzadegan, 2015). This result, in conjunction with our findings, suggests that mathematical education or analytical thinking do not explain the use of correlational thinking. To be able to understand correlational reasoning, which seems to be a robust problem in these tasks, more empirical research regarding the cognitive needs and demands of SF tasks is needed. For example, accumulation is formally known as the integral of its net inflow over time plus the initial stock. In future research one would need to characterize the difficulty and type of mathematical knowledge included in a mathematical test. Also, research would need to establish a causal link in which diverse types of mathematical knowledge are used as interventions and their effects tested experimentally. In our studies, mathematical scores and SF accuracy are only correlated. For example, would emphasis on the concepts of nonlinearities and accumulation improve performance in the SF tasks? Classroom interventions and formal training in system dynamics suggest that educational interventions are possible (Sterman, 2010) but knowledge of the cognitive factors and cognitive demands of the SF tasks is needed (Gonzalez and Wong, 2012; Fischer and Gonzalez, 2015; Lakeh and Ghaffarzadegan, 2015).

Acknowledgements

This research is supported by the National Science Foundation Award, number 1154012, to Cleotilde Gonzalez. The authors would like to thank Hau-yu Wong and Fei Lu for their help and support in data collection and for providing insightful comments. We also thank Professor Xiaorong Liu for her help in data collection in the Chinese study.

DOI: 10.1002/sdr

Biographies

L. Qi is an assistant professor at the Second Military Medical University. He obtained a Ph.D. in Social Medicine and Public Health Management from the same university in 2014.

C. 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 Experimental investigations of human dynamic choice, and computational representations of cognitive processes of dynamic decision making.

References

2013 Shanghai Mathematics Test. 2014. Available: http://shiti.edu.sina.com.cn/paper/6/15/41506/c_p.php [1 March 2014].

Abdel-Hamid T, Ankel F, Battle-Fisher M, Gibson B, Gonzalez-Parra G, Jalali 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.

Beaton AE. 1996. *Mathematics Achievement in the Middle School Years. In IEA's Third International Mathematics and Science Study (TIMSS)*. ERIC: Washington, DC.

Behr MJ, Harel G, Post T, Lesh R. 1992. Rational number, ratio, and proportion. In *Handbook of Research on Mathematics Teaching and Learning*, Grouws DA (ed) ed. Macmillan: New York; 296–333.

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

Brink J, Streefland L. 1979. Young children (6–8): ratio and proportion. *Educational Studies in Mathematics* **10**(4): 403–420.

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

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

Casler K, Bickel L, Hackett E. 2013. Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior* **29**(6): 2156–2160.

Cronin MA, Gonzalez C. 2007. Understanding the building blocks of dynamic systems. *System Dynamics Review* **23**(1): 1–17.

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**(1): 116–130.

De Bock D, van Dooren W, Janssens D, Verschaffel L. 2007. *The Illusion of Linearity: From Analysis to Improvement*, **41**. Springer: Berlin.

Diehl E, Sterman JD. 1995. Effects of feedback complexity on dynamic decision making. *Organizational Behavior and Human Decision Processes* **62**(2): 198–215.

DOI: 10.1002/sdr

Dutt V, Gonzalez C. 2012. Why do we want to delay actions on climate change? Effects of probability and timing of climate consequences. *Journal of Behavioral Decision Making* **25**(2): 154–164.

Fan L, Zhu Y. 2004. How have Chinese students performed in mathematics? A perspective from large-scale international mathematics comparisons. In *How Chinese learn mathematics: Perspectives from insiders*, Fan L, Wong N-Y, Cai J, Li S (eds), **1** ed by. World Scientific: Singapore; 3–26.

Fischer H, Gonzalez C. 2015. Making sense of dynamic systems: how our understanding of stocks and flows depends on a global perspective. *Cognitive Science* 1–17. DOI: 10.1111/cogs.12239

Gonzalez C. 2004. Learning to make decisions in dynamic environments: effects of time constraints and cognitive abilities. *Human Factors* **46**(3): 449–460.

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

Gonzalez C, Vanyukov P, Martin MK. 2005. The use of microworlds to study dynamic decision making. *Computers in Human Behavior* **21**(2): 273–286.

Grow-Maienza J, Hahn D-D, Joo C-A. 2001. Mathematics instruction in Korean primary schools: structures, processes, and a linguistic analysis of questioning. *Journal of Educational Psychology* **93**(2): 363–376.

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

Leung K. 2006. Mathematics education in East Asia and the West: does culture matter?. *In Mathematics Education in Different Cultural Traditions: A Comparative Study of East Asia and the West*, Leung FS, Graf K-D, Lopez-Real F (eds) eds, **9**Vol.. Springer: New York; 21–46.

Mullis IV, Martin MO, Beaton AE, Gonzalez EJ, Kelly DL, Smith TA. 1997. *Mathematics Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: TIMSS International Study Center.

Narayanan A, Moritz BB. 2015. Decision making and cognition in multi-echelon supply chains: an experimental study. *Production and Operations Management* **24**(8): 1216–1234.

Newell B, Kary A, Moore C, Gonzalez C. 2015. Managing the budget: stock–flow reasoning and the CO_2 accumulation problem. *Topics in Cognitive Science* (in press).

Paich M, Sterman JD. 1993. Boom, bust, and failures to learn in experimental markets. *Management Science* **39**(12): 1439–1458.

Sterman JD. 1989a. Misperceptions of feedback in dynamic decision making. *Organizational Behavior and Human Decision Processes* **43**(3): 301–335.

Sterman JD. 1989b. Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment. *Management Science* **35**(3): 321–339.

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.

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

Stevenson HW, Lee S-Y, Stigler JW. 1986. Mathematics achievement of Chinese, Japanese, and American children. *Science* **231**(4739): 693–699.

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

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

Weinhardt JM, Hendijani R, Harman JL, Steel P, Gonzalez C. 2015. How analytic reasoning style and global thinking relate to understanding stocks and flows. *Journal of Operations Management*. http://dx.doi.org/10.1016/j.jom.2015.07.003.

DOI: 10.1002/sdr