Organizational Behavior and Human Decision Processes xxx (2008) xxx–xxx

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

journal homepage: www.elsevier.com/locate/obhdp

Why don't well-educated adults understand accumulation? A challenge to researchers, educators, and citizens

Matthew A. Cronin [a,*], Cleotilde Gonzalez [b], John D. Sterman [c]

[a] School of Management, George Mason University, Mail Stop 5F5, Fairfax, VA 22030-4444, USA
[b] Dynamic Decision Making Laboratory, Social and Decision Sciences Department, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, USA
[c] System Dynamics Group, MIT Sloan School of Management, 30 Wadsworth Street, E53-351, Cambridge, MA 02141, USA

ARTICLE INFO

Article history:
Received 10 December 2006
Accepted 31 March 2008
Available online xxxx

Accepted by Robyn Dawes

Keywords:
Dynamic decision making
Cognitive capacity
Accumulation
Stocks and flows
System dynamics

ABSTRACT

Accumulation is a fundamental process in dynamic systems: inventory accumulates production less shipments; the national debt accumulates the federal deficit. Effective decision making in such systems requires an understanding of the relationship between stocks and the flows that alter them. However, highly educated people are often unable to infer the behavior of simple stock–flow systems. In a series of experiments we demonstrate that poor understanding of accumulation, termed *stock–flow failure*, is a fundamental reasoning error. Persistent poor performance is not attributable to an inability to interpret graphs, lack of contextual knowledge, motivation, or cognitive capacity. Rather, stock–flow failure is a robust phenomenon that appears to be rooted in failure to appreciate the most basic principles of accumulation, leading to the use of inappropriate heuristics. We show that many people, including highly educated individuals with strong technical training, use what we term the "correlation heuristic", erroneously assuming that the behavior of a stock matches the pattern of its flows. We discuss the origins of stock–flow failure and implications for management and education.

Understanding and managing stocks and flows—that is, resources that accumulate or deplete and the flows that alter them—is a fundamental process in society, business, and personal life. At the macroeconomic level, for example, exploration increases known petroleum reserves, while oil production reduces the stock of oil remaining for the future. In turn, petroleum combustion increases the stock of carbon dioxide in the atmosphere and contributes to global warming. At the organizational level, firms' capabilities and competitive advantages arise from the accumulation of resources and knowledge. Firms must manage their cash flows to maintain adequate stocks of working capital, and production must be adjusted as sales vary to maintain sufficient inventory. Individuals, too, face similar stock management challenges: we manage our bank accounts (stock of funds) to maintain a reasonable balance as our incomes (inflows) and expenses (outflows) vary, and we struggle to maintain a healthy weight by managing the inflow and outflow of calories through diet and exercise. Accumulation is a pervasive process in everyday life, and arises at every temporal, spatial and organizational scale.

All stock–flow systems share the same underlying structure. The resource level (stock) accumulates the inflows to it less the outflows from it.[1] Although the relationship between stocks and flows is a fundamental concept of calculus, knowledge of calculus is not necessary to understand the behavior of stocks and flows. Any stock can be thought of as the amount of water in a tub. The water level accumulates the flow of water into the tub (the inflow)

* Corresponding author.
E-mail addresses: mcronin@gmu.edu (M.A. Cronin), conzalez@andrew.cmu.edu (C. Gonzalez), jsterman@mit.edu (J.D. Sterman).

[1] Formally, consider any stock, S, with inflow I and outflow O. The stock at any time T, S_T, is the integral of its net inflow over time, plus the quantity in the stock at the initial time, t_0, S_{t0}. The net inflow to the stock is the difference between inflow and outflow, $I - O$:

$$S_T = \int_{t_0}^{T} \text{Net inflow}\, dt + S_{t_0} = \int_{t_0}^{T} (I - O)\, dt + S_{t_0}$$

Equivalently, the rate of change of the stock is the net inflow:

$$\frac{dS}{dt} = \text{Net inflow} = I - O$$

Note the units of measure: the stock is measured in units, e.g., liters of water in a tub, dollars in an account, or people in a store, while the inflow, outflow, and net rate of change in the stock are measured in units/time period, e.g., litres/sec, $/month, or people/min.

0749-5978/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.obhdp.2008.03.003

less the flow exiting through the drain (the outflow). The rate of change in the water level is the net flow, given by the difference between the inflow and outflow. As everyday experience suggests, the water level rises only when the inflow exceeds the outflow, falls only when the outflow exceeds the inflow, and remains the same only when the inflow equals the outflow.

Stock–flow problems, even simple ones, are unintuitive and difficult, even for highly educated people with strong mathematics backgrounds, including calculus (Booth Sweeney & Sterman, 2000; Cronin & Gonzalez, 2007; Sterman & Booth Sweeney, 2002). For example, Booth Sweeney and Sterman (2000) presented graduate students at an elite university with a picture of a bathtub and graphs showing the inflow and outflow of water, then asked them to sketch the trajectory of the stock of water in the tub. Although the patterns were simple, fewer than half responded correctly. We denote such difficulties *stock–flow (SF) failure*.

This paper investigates the sources of and psychological processes involved in SF failure. We address many commonly cited reasons for poor performance in dynamic decision-making studies and stock–flow contexts (e.g., Atkins, Wood, & Rutgers, 2002; Booth Sweeney & Sterman, 2000; Jensen & Brehmer, 2003; Sterman & Booth Sweeney, 2007). Experiments 1–4 test whether information displays (Atkins et al., 2002), the cognitive burden of required calculations (Roch, Lane, & Samuelson, 2000), inadequate motivation, unfamiliar task context, or the inability to interpret or construct graphs (Ossimitz, 2002) contribute to SF failure. The experiments demonstrate the persistence of SF failure even when the task is simple, participants are motivated, contexts are familiar, information displays are varied and participants are highly educated and able to read graphs. The results suggest that SF failure arises from a deeper and more robust difficulty, analogous to the persistent problems people have in probabilistic judgment and decision making (Dawes, 1988, 1998; Kahneman & Tversky, 1972).

It appears that many people have difficulty applying the principles of accumulation correctly, failing to grasp that the quantity of any stock, such as the level of water in a tub, rises (falls) when the inflow exceeds (is less than) the outflow. Rather, it appears that people often use intuitively appealing but erroneous heuristics such as assuming that the output of a system is positively correlated with its inputs. That is, people assume that the output (the stock) should "look like" the input (the flow or net flow). We denote such behavior the *correlation heuristic*.

Correlational reasoning can be useful and adaptive (e.g., illness is highly correlated with the consumption of certain mushrooms; the rustling of leaves in the underbrush often heralds a stalking predator). Further, in some cases, by accident, the correlation heuristic yields the correct response in a stock–flow situation. Specifically, when the net inflow to a stock is positive and growing exponentially, the stock will also grow exponentially.[2] Many macroeconomic variables do have strong exponential trends, for example, GDP, the federal deficit, and greenhouse gas emissions. However, the fact that the correlation heuristic works when the net flow is exponential is an accident. The correlation heuristic fails with any departure from a pure exponential growth trend for the net flow.

These failures of the correlation heuristic are highly consequential. For example, the US federal deficit and national debt have both risen dramatically in the past half-century, and they are highly correlated ($r = .80$ for annual data from 1950 to 2005, $p < .001$). However, because the national debt is a stock that accumulates the deficit, it will continue to rise even if the deficit falls—the debt can fall only if the government runs a surplus. Similarly, anthropogenic greenhouse gas (GHG) emissions are now roughly double the rate at which they are removed from the atmosphere by natural processes (Houghton et al., 2001; IPCC, 2007). Therefore, atmospheric greenhouse gas concentrations will continue to rise even if emissions fall, until emissions fall to the rate at which GHGs are removed from the atmosphere. However, experiments show that the vast majority of adults believe GHG concentrations follow the same pattern as emissions, leading them to conclude, erroneously, that atmospheric GHG concentrations can be stabilized even as emissions into the atmosphere continuously exceed removal from it (Sterman & Booth Sweeney 2007). Such beliefs are analogous to the assertion that a bathtub continuously filled faster than it drains will never overflow. They violate conservation of mass and lead to the erroneous conclusion that climate change can be mitigated simply by slowing the growth of emissions. Correlational reasoning can lead to erroneous judgments in situations with important public policy implications.

The paper proceeds as follows: We describe prior work in the context of a simple stock–flow problem. Experiments 1–4 test alternative explanations for SF failure including the cognitive burden of the task, information display, task context, motivation and feedback, and priming of prior stock–flow knowledge. Close analysis of the data suggest participants use correlational reasoning across these contexts. In the final experiment we test the correlation heuristic directly, showing that the large majority assume the stock is correlated with the inflow even in extremely simple situations. We consider limitations and extensions, discuss the managerial and educational implications of the results, and offer suggestions for research to create interventions that may overcome SF failure.

A simple stock and flow problem

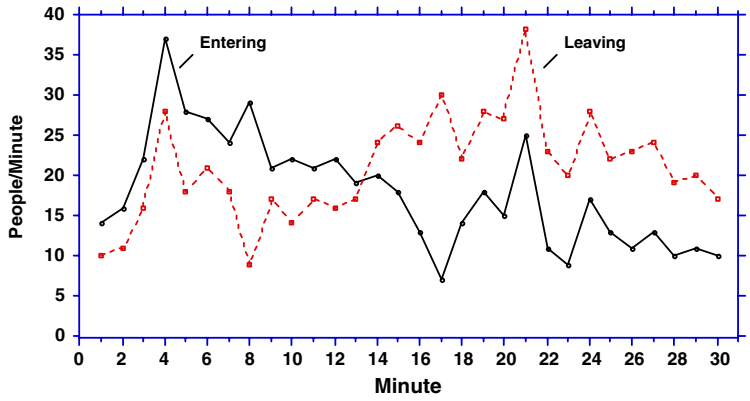
Prior work in dynamic decision making suggests that people have great difficulty understanding and managing systems with high levels of dynamic complexity. Dynamic complexity arises from the presence of multiple feedback processes, time delays, nonlinearities, and accumulations (Sterman, 2002). Furthermore, learning in dynamic systems is often slow and weak, even with repeated trials, unlimited time, and performance incentives (Diehl & Sterman, 1995; Kleinmuntz & Schkade, 1993; Sterman, 1989a, 1989b). Many of these studies involved tasks of great complexity, and poor performance was often ascribed to the large number of entities and interactions, feedback delays, and information overload (Brehmer, 1990, 1995; Gonzalez, 2005a; Kleinmuntz, 1985; Omodei & Wearing, 1995).

More recent work, however, has shown that people make persistent mistakes even in the simplest dynamic systems, including systems consisting of one stock, one inflow, and one outflow, with no feedback processes, time delays, or nonlinearities (e.g., Booth Sweeney & Sterman, 2000; Cronin & Gonzalez, 2007; Sterman & Booth Sweeney, 2007). To illustrate, the "department store" task (Sterman, 2002) presents participants with a graph showing the number of people entering and leaving a department store each minute over a 30-min interval (Fig. 1). The system involves a single stock (the number of people in the store) with one inflow (people entering) and one outflow (people leaving). There are no feedbacks, time delays, nonlinearities, or other elements of dynamic complexity that proved to be difficult in prior research. Participants are asked four questions. The first two—"When did the most people enter the store? When did the most people leave the store"—test whether participants can read the graph and correctly distinguish

[2] The only function whose integral is the function itself (that is, where $\int f(x)dt = f(x)$, so that the net flow and stock are perfectly correlated) is the exponential function. Thus, when the net flow into a stock is growing exponentially, Net Flow = $\exp(gt)$, the stock follows an exponential path, Stock = $\exp(gt)/g$.

Please cite this article in press as: Cronin, M. A. et al., Why don't well-educated adults understand accumulation? A challenge ..., *Organizational Behavior and Human Decision Processes* (2008), doi:10.1016/j.obhdp.2008.03.003

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



Please answer the following questions.

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

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

Minute _____ ❑ Can't be determined

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

Minute _____ ❑ Can't be determined

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

Minute _____ ❑ Can't be determined

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

Minute _____ ❑ Can't be determined

Fig. 1. Department store task.

between inflow and outflow. The next questions—"When were the most people in the store? When were the fewest people in the store?"—test whether participants can infer the behavior of the stock from the behavior of the flows.

To answer one could manually accumulate the stock by finding the net inflow at each point of time and keep a running tally of the number of people in the store ($S_t = S_{t-1} + I_t - O_t$). This method, however, is tedious, error prone, and unnecessary. One need only understand that the number of people in the store rises when the flow of people entering is greater than the flow of people leaving (and vice versa), then note that the number entering is greater than the number exiting through time 13 and less thereafter. Therefore, one can see—without any calculations—that the most people are in the store when the two curves cross (Minute 13). Furthermore, because the number of people in the store rises through Minute 13 and falls thereafter, the fewest people are in the store either at the beginning or the end of the 30 min. To determine which, participants must judge whether more people (net) enter up to Minute 13 than leave afterward. Once again, calculation is unnecessary: One can simply judge whether the area between the rate of entering and the rate of leaving up to Minute 13 is greater or smaller than the area between the two curves from Minute 14 on. The area between the curves from Minute 14 on is clearly larger, so the fewest people are in the store at the end of the 30 min.[3]

Method

A total of 173 students enrolled in a graduate course in systems thinking and simulation at the MIT Sloan School of Management were given the department store task in Fig. 1. Participants were primarily MBA students and graduate students from other MIT departments or from Harvard University. The mean age was 29 (range 21–46) and 78% were male. All had taken calculus, and most had strong mathematics training: 71% had a degree in science, technology, engineering, or mathematics (STEM); 28% had a degree in the social sciences, primarily economics. Fully 40% had a prior graduate degree, most in technical fields. Students did the task in

[3] It might be objected that judging the areas of the irregular shapes defined by the difference between inflow and outflow in Fig. 1 is difficult. However, the task was carefully designed to make the determination of the area simple. The area of the region in which outflow exceeds inflow (after $t = 13$) was constructed to be twice as large as the area in which inflow exceeds outflow (prior to $t = 13$). To test whether people can determine which area is larger, a convenience sample consisting of 12 members of the support staff from the MIT Sloan School of Management were asked which area was greater; all correctly identified the larger area.

Please cite this article in press as: Cronin, M. A. et al., Why don't well-educated adults understand accumulation? A challenge ..., *Organizational Behavior and Human Decision Processes* (2008), doi:10.1016/j.obhdp.2008.03.003

Table 1
Results of the baseline department store task

	Most entering		Most leaving		Most in store		Fewest in store	
	N	%	*N*	%	*N*	%	*N*	%
Max entering $t = 4$	**166**	**96.0**	0	0	6	3.5	1	0.6
Max leaving $t = 21$	2	1.2	**164**	**94.8**	1	0.6	3	1.7
Max in store $t = 13$	0	0	0	0	**76**	**43.9**	4	2.3
Fewest in store $t = 30$	0	0	0	0	1	0.6	**54**	**31.2**
Max net inflow $t = 8$	4	2.3	0	0	50	28.9	0	0
Max Net Outflow $t = 17$	0	0	6	3.5	6	3.5	51	29.5
Initial in store $t = 1$	0	0	0	0	0	0	12	6.9
Cannot be determined	0	0	0	0	29	16.8	43	24.9
Other	1	0.6	2	1.2	2	1.2	2	1.2
No answer	0	0	1	0.6	2	1.2	3	1.7

Rows show the responses with the time point specified in the first column (±1 min to account for possible participant error in reading time-axis values). Columns correspond to the number (Total $N = 173$) and percent of people selecting a particular answer for each question in Fig. 1. For example, 166 selected $t = 3$, 4, or 5 min to answer "During which minute did the most people enter the store?" (the correct answer is 4). There were no significant differences in response frequencies by question order so the two question order treatments are collapsed in the table (see text). Bold entries highlight correct responses. Underlined cells show those incorrectly specifying the maximum net inflow/net outflow instead of maximum/minimum in the store.

class at the beginning of the semester. Students were given approximately 10 min. Participation was voluntary. Students were informed that the results would not be graded but illustrated important concepts they were about to study and would be used anonymously in this research. To test for order effects, half the participants were randomly selected to receive the questions about the flows (1 and 2) first (Order O1), and half received the two questions about the stock (3 and 4) first (Order O2).

Results

Table 1 summarizes the results of the department store task.[4] Question order made no difference (Fisher's exact test for differences in the proportion correct on each of the four questions by presentation order yielded $p = .44$, .17, 1.0, and 1.0, respectively). Hence, we pool O1 and O2 in Table 1 and the results presented here. The vast majority of participants correctly identified when the most people entered and left the store (96% and 95% for Questions 1 and 2, respectively). However, few were able to answer the stock–flow questions correctly (44% and 31% for Questions 3 and 4, respectively). About 17% indicated that it is not possible to determine when the most people were in the store, and 25% said that it is not possible to determine when the fewest people were in the store. More importantly, 29% incorrectly indicated that the most are in the store when the net inflow is greatest ($t = 8$) and 30% incorrectly conclude that the fewest are in the store when the net outflow is greatest ($t = 17$). These responses, accounting for far more of the erroneous choices than any other, reveal a fundamental confusion about the relationship between stocks and flows.

Our next experiments seek to illuminate why performance is so poor. In each case, participants were given a problem involving one stock, one inflow, and one outflow. Participants were asked to answer the four questions presented in the baseline experiment. Because the baseline condition revealed no order effect, the questions were always presented in the same order (most entering, most leaving, most in store, fewest in store). Unless otherwise noted, participants received as much time as needed (rarely did they need longer than 10 min) and were offered no performance-based incentive (except in Exp. 3).

[4] Answers to all questions were considered correct if they were within 1 min of the correct response, that is, responses of 12, 13, or 14 were coded as correct responses to the question "During which minute did the most people enter the store?" These tolerances count as correct those who understood the concepts but might have misread the time-axis values, favoring high performance.

Experiment 1: Cognitive burden and data display

Limited cognitive capacity is a commonly cited explanation for poor problem solving performance in general (Simon, 1979) and in dynamic decision making specifically (Atkins et al., 2002; Roch et al., 2000). To calculate the number of people in the store each minute, participants must read the graph to estimate the numerical value of the flow of people entering and leaving the store, then subtract the outflow from the inflow to compute the net flow, and finally, add the net flow to the running tally of the stock stored in memory. The baseline task in Fig. 1 presents 60 data points (inflow and outflow data for 30 min), perhaps overwhelming participants' cognitive capacity and working memory. Therefore, we created a simpler version (Fig. 2A). The pattern is similar to that of the baseline condition but presents data for only 12 min, and the value of the flows never exceeds 15 people per minute (compared to nearly 40 people per minute in the baseline task). We retained the key features of the baseline task: The number of people in the store rises, peaks, and falls. The most people are in the store at $t = 7$, and the fewest are in the store at the end ($t = 12$). The number leaving after the population peaks is twice the number entering before the peak. If cognitive capacity is a source of error, then:

$H_{1.1}$: Performance will improve in simpler versions of the task with fewer data points.

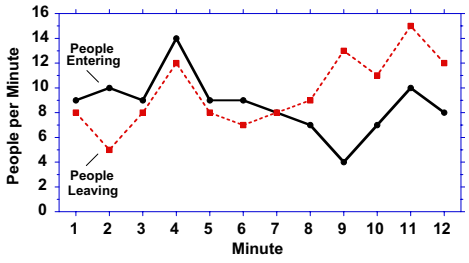
If, on the other hand, the difficulty arises from a weak understanding of the concept of accumulation, then performance in the simpler version should not improve.

Another common explanation for poor problem solving is confusing information display. Poor ability to interpret and construct graphs among both students and adults is well documented in the mathematics education literature and related fields (e.g., Berry & Nyman, 2003; Carlson, Jacobs, Coe, Larsen, & Hsu, 2002; Gattis, 2002, 2004; Gattis & Holyoak, 1996; Paulos, 1988; Tufte 1983, 1990). Such studies suggest that poor performance may not reflect poor understanding of the concept of accumulation but rather difficulty in extracting the information needed to answer the questions from graphical displays. To test this possibility, we created three alternative visual presentations of the data in Fig. 2A: a bar graph, a table, and a textual presentation (see Figs. 2B–D).

The high performance on Questions 1 and 2 suggests that participants in the baseline condition (MIT and Harvard graduate students) were able to read the graphs. Nevertheless, perhaps the use of a graphical display makes it difficult to appreciate the accumulation of people in the store. If the ability to interpret graphs is the

Please cite this article in press as: Cronin, M. A. et al., Why don't well-educated adults understand accumulation? A challenge ..., *Organizational Behavior and Human Decision Processes* (2008), doi:10.1016/j.obhdp.2008.03.003

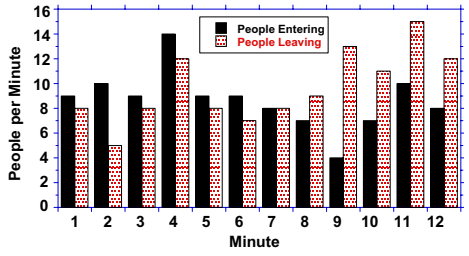
A. Line graph



C. Table

Minute	People Entering	People Leaving
1	9	8
2	10	5
3	9	8
4	14	12
5	9	8
6	9	7
7	8	8
8	7	9
9	4	13
10	7	11
11	10	15
12	8	12

B. Bar Graph



D. Text

In the first minute, 9 people enter and 8 leave. In the second minute, 10 people enter and 5 leave. In the third minute, 9 people enter and 8 leave. In the fourth minute, 14 people enter and 12 leave. In the fifth minute, 9 people enter and 8 leave. In the sixth minute, 9 people enter and 7 leave. In the seventh minute, 8 people enter and 8 leave. In the eighth minute, 7 people enter and 9 leave. In the ninth minute, 4 people enter and 13 leave. In the tenth minute, 7 people enter and 11 leave. In the eleventh minute, 10 people enter and 15 leave. In the twelfth minute, 8 people enter and 12 leave.

Fig. 2. Experiment 1: Visual isomorphs for the simpler department store task.

source of the difficulty, then an alternate data presentation mode, such as a table or text, should improve performance on the stock–flow questions. Similarly, if people attempt to calculate the number of people in the store by the running total method, then graphical displays will be more difficult than tabular or textual displays because one must first estimate the numerical values of the inflow and outflow from the graph. If the problem is difficulty interpreting graphical displays, then

$H_{1.2}$: Performance will improve if the data are presented in tabular or textual form.

A more nuanced theory suggests that the *form* of the information display leads people to misinterpret the problem. Atkins et al. (2002) review the literature showing how display formats affect judgment and decision making, and, following Wickens and Carswell (1995), suggest that information displays can be more or less compatible with the demands of the task, thus affecting performance, even if the information they encode is the same. Vicente and Rasmussen (1992) show that interface format can lead to decision error and call for "ecological interface design" where display type and content are matched to the decision context; Vicente (1996) provides an example in a dynamic decision task.

The information in Figs. 1 and 2A is presented using a line graph. Line graphs are often used to represent continuously varying quantities, such as water flowing into a tub. Here, however, the data points represent the total number of discrete individuals who entered or left over the course of each minute, not the instantaneous rate at each moment. The continuous flow metaphor suggested by the line graph may conflict with participants' conception of the discrete event of a person entering or leaving a store. Bar graphs are more commonly used to represent totals over some finite period and may help people recognize and understand the relationship between the flows and the stock. If so,

$H_{1.3}$: People will be more successful at judging the behavior of stocks and flows when discrete quantities are represented with discrete features (bars instead of lines).

Comparing the table to a textual presentation (Fig. 2C vs. D), participants should find a table more helpful than text because the numbers are already aligned, reducing the cognitive burden involved in finding the inflow and outflow values and calculating the net flow. Hence,

$H_{1.4}$: Presenting the data in a table will improve performance compared to text.

Method

Participants ($N = 271$) were students enrolled in a subsequent term of the same graduate course at the MIT Sloan School of Management used in the baseline experiment. The average age was 29 (range 18–38), and 69% were male. Of the participants, 55% were trained in STEM, 38% were trained in economics or other social science and 29% held prior advanced degrees. Participants were demographically similar to those who participated in the baseline

Please cite this article in press as: Cronin, M. A. et al., Why don't well-educated adults understand accumulation? A challenge ..., *Organizational Behavior and Human Decision Processes* (2008), doi:10.1016/j.obhdp.2008.03.003

experiment in gender, age, prior education, and prior advanced degrees. Participants were randomly assigned to one of the four data presentation modes shown in Fig. 2. As in the baseline, responses were considered correct if they were within 1 min of the correct answer.

Results

We first consider differences across data display modes. Table 2, Block A compares performance in the two graphic conditions (line and bar) to the two non-graphic conditions (table and text). Contrary to the hypothesis that SF failure results from participants' inability to interpret graphs, graphic presentation appears to reduce errors in data interpretation. Performance on Questions 1 and 2 was significantly higher for the graphic displays ($p = .015$ and .001, respectively). Contrary to Hypothesis 1.2, performance on the two stock–flow questions was no better in the non-graphic conditions. There is no support for the hypothesis that difficulties in interpreting graphs are responsible for the participants' poor understanding of accumulation.

Table 2, Block B compares the line and bar graph representations. People's ability to determine when the most people enter and leave the store is not statistically different across the line and bar graph conditions ($p = .72$ and .13, respectively). Performance on Question 3 (when are the most in the store?) was slightly better in the bar graph condition, but the difference is only marginally significant, $p = .07$, and there was no difference between the bar and line graph condition on Question 4 (when are the fewest in the store?), $p = .16$. At best, there is only weak support for Hypothesis 1.3: Participants did not appear to be confused by the presentation of the data in the line graph format, which suggests continuous flows, compared to the bar graph format, which suggests discrete flows. There also was no difference between the tabular and textual presentations on any of the questions (Block C), providing no support for $H_{1.4}$.

Turning now to the issue of cognitive overload, Block D in Table 2 compares the performance of participants who received the simple line graph condition (Fig. 2A) to the baseline experiment (Fig. 1).[5] Results are similar to the baseline. Performance on Questions 1 and 2 was high (participants correctly interpret the graphs), but performance on the stock–flow questions was poor. Individuals receiving the simpler version did no better than those receiving the baseline condition. Like the baseline, many participants in the line graph condition indicated that the answer to the stock questions could not be determined (21% and 27% for Questions 3 and 4, respectively, compared to 17% and 25% in the baseline; $p = .44$ and .86). Thus, Hypothesis 1.1 is not supported: The simpler version of the task with far fewer data points did not improve performance on the stock–flow questions.

One may argue that, although the simplified graph, with 12 rather than 30 min of data, reduces the mental burden of the task, it still overwhelms participants' cognitive capacity. However, even simpler versions with still fewer data points and even simpler patterns do not improve performance, as in Experiment 4 below and in Cronin and Gonzalez (2007). Overall, the results of Exp. 1 do not support the hypotheses that cognitive capacity, the ability to interpret graphs, or the mode of information display cause poor performance in stock–flow systems.

[5] Because the baseline task in Fig. 1 and the simpler tasks in Fig. 2 were administered to students in successive years, it is possible that unmeasured sources of variation could confound the interpretation of the results. However, the tasks were given to each group by the same instructor (JS), in the same course, at the same point in the semester, in the same classroom, at the same time of day, and with the same instructions and time for completion. The two groups were demographically similar. Nevertheless, we alert the reader to the possibility that the results could reflect unmeasured differences across the two groups.

Table 2
Experiment 1: Success rates between visual isomorphs

		Question 1: Most entering?	Question 2: Most leaving?	Question 3: Most in store?	Question 4: Fewest in store?
Overall success rate ($N = 264$)		89%	83%	56%	46%
A	Graph (both line and bar, $N = 127$)	94%	91%	61%	48%
	No graph (both text and table, $N = 137$)	85%	76%	51%	44%
	Exact test (p)	0.015	0.001	0.137	0.537
B	Line graph ($N = 63$)	94%	87%	52%	41%
	Bar graph ($N = 64$)	95%	95%	69%	55%
	Exact test (p)	0.718	0.127	0.071	0.157
C	Text ($N = 59$)	86%	75%	47%	42%
	Table ($N = 78$)	83%	77%	54%	45%
	Exact test (p)	0.811	0.841	0.493	0.862
D	Baseline ($N = 173$)	96%	95%	44%	31%
	Line graph ($N = 63$)	94%	87%	52%	41%
	Exact test (p)	0.490	0.083	0.302	0.166

Experiment 2: Task context

Framing conditions choice. For example, people find the classic Wason (1960) confirmation bias task difficult when presented as an abstract test of a logical rule ("if a card has a vowel on one side then it must have an even number on the other"), but do much better when the cover story is a familiar everyday context such as "if an employee gets a day off during the week, then that employee must have worked on the weekend" (Gigerenzer & Hug, 1992). Such results suggest that reasoning is often domain-specific and adapted to specific contexts (but see also Almor & Sloman, 2000). Perhaps people understand the principles of accumulation but fail to recognize the stock and flow structure of the situation because the task context or cover story does not activate their latent stock–flow knowledge or experience with other accumulation processes. If so, even those who have studied calculus may not recognize the stock–flow structure of the department store context and therefore may not apply their knowledge of accumulation. Therefore, we created two additional cover stories for the original baseline task outlined in Fig. 1. In the *tub condition,* the data represent the flow of water into and out of a bathtub, and the stock is the quantity of water in the tub. In the *cars condition,* the data represent the velocities of two cars traveling in the same direction; the stock is the distance between them.

We hypothesize that the likelihood of activating and applying the stock–flow schema depends on the salience and familiarity of the accumulation process in the task context. Most people have more experience and familiarity with bathtubs and driving than with the flow of people into and out of stores. Accumulation is the purpose and focus of attention when filling a container, and people directly observe the flows and water level. Similarly, monitoring the distance between vehicles is a central task in driving, and the speed and distance between cars are directly observable. In contrast, the flows of people into and out of a store and the population of people within it are usually neither observable nor important in everyday experience. If salience and familiarity are important in activating latent stock–flow knowledge, then performance in the tub and cars conditions will be better than in the store condition.

Please cite this article in press as: Cronin, M. A. et al., Why don't well-educated adults understand accumulation? A challenge ..., *Organizational Behavior and Human Decision Processes* (2008), doi:10.1016/j.obhdp.2008.03.003

The activation of latent stock–flow knowledge may also depend on whether the cover story involves discrete events or continuous flows. Common examples of accumulation used in high school mathematics and physics classes, for example, involve continuous quantities and flows, such as water filling a tank or velocity accumulating into distance traveled. The store context, however, involves discrete individuals entering and leaving at particular moments, which may prevent participants from recognizing the stock–flow structure. If people's intuitive understanding of accumulation is grounded in schemata that are based on continuous flows, then the store context, with its discrete individuals, may not activate their latent stock–flow knowledge. Hence, if either familiarity or continuous flows are important triggering people's knowledge of accumulation,

H_2: Performance in the tub and cars conditions will be better than in the store condition.

Method

We recruited 47 undergraduate students from Carnegie Mellon University who participated voluntarily and received $5.00 compensation for their time. The average age was 25. No other demographic information was collected in this case. Participants were randomly assigned to the tub, cars, or store condition.

Results

As in prior experiments, this population showed excellent ability to read the graph, while performance on the stock–flow questions was poor (Table 3), with 96% and 94% correctly answering Questions 1 and 2, respectively, but only 28% and 26% answering Questions 3 and 4 correctly. There were no statistically significant differences in performance on the stock questions across the different task contexts. Hypothesis 2 is not supported: The more familiar tub and driving contexts, with continuous rather than discrete flows, do not improve performance. Even if people find it difficult to recognize the stock–flow structure in the cars context, water flowing into a tub is a canonical example of a stock–flow system. The fact that performance did not differ across these conditions suggests that people have difficulty applying the principles of accumulation even in settings where the presence of accumulation is obvious.

Table 3
Experiment 2: Success rates across cover stories

		Question 1: Largest inflow	Question 2: Largest outflow	Question 3: Stock most full	Question 4: Stock most empty
Overall success rates ($N = 47$)		96%	94%	28%	26%
A	Store ($N = 18$)	100%	100%	22%	17%
	Cars ($N = 16$)	100%	100%	38%	31%
	Exact test (p)	1.000	1.000	.457	.429
B	Store ($N = 18$)	100%	100%	22%	17%
	Tub ($N = 13$)	85%	77%	23%	31%
	Exact test (p)	.168	.064	1.000	.413
C	Tub + cars ($N = 29$)	93%	90%	31%	31%
	Store ($N = 18$)	100%	100%	22%	17%
	Exact test (p)	.517	.276	.739	.324
D	Tub ($N = 13$)	85%	77%	23%	31%
	Cars ($N = 16$)	100%	100%	38%	31%
	Exact test p=	.192	.078	.454	1.000

Experiment 3: Motivation and feedback

Another potential explanation for SF failure is that people lack the motivation to think deeply about their answers. In the baseline task (Fig. 1) and in some prior research (Booth Sweeney & Sterman, 2000), no incentives were offered for performance, perhaps reducing motivation and effort. The role of incentives in decision making is complex: incentives sometimes improve performance, sometimes have no impact, and sometimes actually worsen performance (Camerer & Hogarth, 1999 provide a review). However, people who do not have a reason to think hard about a problem tend to rely on simple heuristics instead of solving problems analytically or challenging commonsense frames (Petty & Wegener, 1998). Incentives have also been shown to increase motivation to solve problems analytically (Petty, Wegener, & Fabrigar, 1997).

With low motivation, people may devote insufficient cognitive effort to the problem and use the correlation heuristic, quickly yielding an answer that appears to be correct. Higher motivation should encourage greater cognitive effort and care in retrieving and applying whatever latent stock–flow knowledge people may possess. Further, for those who do not understand that any stock rises when inflow exceeds outflow and falls when outflow exceeds inflow—which would allow them to answer correctly without calculation—higher motivation should lead more participants to find the correct response by calculating the running total store population, improving performance even if they do not understand the principles of accumulation. Thus,

$H_{3.1}$: High motivation will improve performance on the stock–flow questions.

Low motivation may also lead people to fail to check their answers, resulting in mistakes that might easily be corrected if people received outcome feedback. People often assume that their initial intuition about a problem is correct unless they receive feedback (Klayman & Ha, 1987; Raman, 2002). Therefore, if people actually understand the principles of accumulation but make careless mistakes due to inattention or low motivation, then feedback alerting people to error in their initial judgments should improve performance by inducing greater cognitive effort in subsequent attempts:

$H_{3.2}$: Feedback that alerts participants to mistakes will improve the rate of success on subsequent attempts.

Method

We recruited 69 undergraduate students from the George Mason University School of Management, all of whom participated for course credit. The treatment group ($N = 32$) received the motivation/feedback condition; the control group ($N = 37$) received the no motivation/no feedback condition. The average age of the participants was 23 (range 19–50), and 48% were male. Participants in both conditions received the standard protocol for the task outlined in Fig. 3 and were given up to one hour to complete the task.

In the no motivation/no feedback condition, participants only had to answer the four questions and received no performance-based reward. In the motivation/feedback condition, participants were instructed to bring their papers to the experimenter to find out whether their answers were correct. Incorrect responses were marked wrong, but no other information was provided. The participants returned to their seats with the same graph to correct their response(s). Participants then turned in their sheets to the experimenter and again received feedback, continuing this process until they answered all four questions correctly. Motivation was in-

Please cite this article in press as: Cronin, M. A. et al., Why don't well-educated adults understand accumulation? A challenge ..., *Organizational Behavior and Human Decision Processes* (2008), doi:10.1016/j.obhdp.2008.03.003

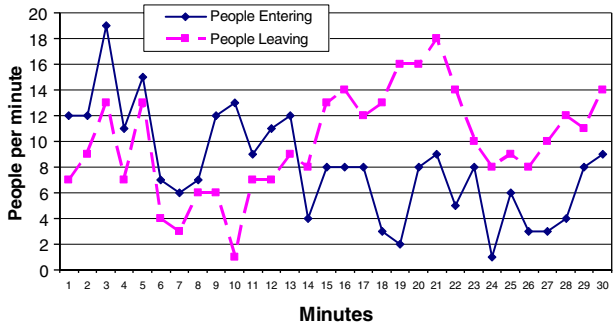


Fig. 3. Experiment 2: Graph used in motivation experiments.

duced by informing participants that they could leave the session once they answered all questions correctly or after one hour, whichever came first. Participants in the prior experiments normally spend less than 10 min answering the four questions (this was also true for the no motivation/no feedback group), so those answering correctly could save the bulk of an hour, motivating them to do well on the first attempt.

The effect of motivation was assessed by comparing the performance of those in the no motivation/no feedback condition to the performance of those in the motivation/feedback condition on the first attempt. The effect of feedback was assessed by comparing the percentage of people in the motivation/feedback condition who answered the question correctly on the first try to the percentage of those who answered correctly on the second try after receiving feedback.

Results

Results for participants' first attempt are similar to prior conditions: nearly all of the participants read the graph correctly on their first attempt (Table 4A), but very few answered the stock–flow questions correctly (16% and 13% for Questions 3 and 4, respectively). Motivation did not significantly improve performance, thus Hypothesis 3.1 is not supported.

If SF failure was simply a careless mistake that could be easily corrected, then performance should have improved quickly upon receiving feedback. Yet while feedback did eventually improve performance, the rate of improvement did not increase across attempts. For example, 5 of 32 participants answered Question 3 correctly the first time (16%). Of the remaining 27 participants, 4 answered correctly on the second attempt (15%; see Table 4B). Hypothesis 3.2 is not supported: There is no statistically significant difference in success rates on the stock–flow questions after the participants received feedback. The cumulative number of participants who answered the stock–flow questions correctly rose slowly and with a decreasing slope (see Table 4A): only 28% and 25% correctly answered Questions 3 and 4, respectively, on the second attempt, and by the sixth attempt, performance reached only 81% and 84%, respectively. There was no further improvement with continued trials; the remaining participants were unable to answer the stock–flow questions by the end of the hour and were dismissed. The mean number of attempts required to answer both stock–flow questions correctly was 4.6.

Table 4B
Experiment 3: Effect of Feedback in the High-Motivation Condition

	Question 3: Most in store?	Question 4: Fewest in store?
Correct on first try	5 of 32 (15.6%)	4 of 32 (12.5%)
Correct on second try	4 of 27 (14.8%)	4 of 28 (14.3%)
Exact test (p)	1.00	1.00

Experiment 4: Priming stock–flow knowledge

In Experiment 4 we take a more direct approach to testing people's ability to understand accumulation. We ask people to calculate a stock from information on its flows in an extremely simple setting (constant flows), asking them to provide the value of the stock every period. Doing so should activate people's latent SF knowledge (if it exists), thus:

H_4: Priming participants to notice the presence and behavior of SF structures will increase performance on subsequent stock–flow tasks.

Method

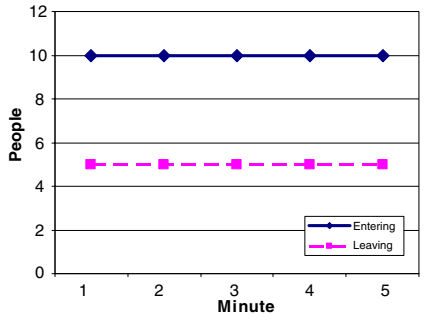
We recruited 37 undergraduate students at George Mason University, all of whom participated for course credit. Their average age was 23 (range 19–44), and 42% were male. Participants were first given a priming task showing a constant inflow of 10 people per minute and a constant outflow of 5 people per minute, over an interval of 5 min (Fig. 4A). Written instructions asked participants to determine how many people are in the store each minute, starting with no one in the store. The explicit direction to record how many are in the store each minute should help participants recognize that the number of people in the store accumulates the inflow less the outflow without explicitly telling people how to do the calculation. The extreme simplicity of the example reduces the cognitive burden of the required calculations. Immediately after completing the priming

Table 4A
Experiment 3: Effect of motivation and feedback on success rates for task 1

	Question 1: Most entering?	Question 2: Most leaving?	Question 3: Most in store?	Question 4: Fewest in store?
No motivation/no feedback condition (n = 37)	100%	86.5%	18.9%	21.6%
Motivation/feedback condition (n = 32): Attempt 1	93.8%	96.9%	15.6%	12.5%
Exact test, p	.211	.205	.761	.359
Attempt 2	100%	100%	28.1%	25.0%
Attempt 3			56.3%	50.0%
Attempt 4			65.6%	62.5%
Attempt 5			68.8%	71.9%
Attempt 6			71.9%	81.3%
Attempt 7			81.3%	84.4%
Attempt 8			81.3%	84.4%
Attempt 9			81.3%	84.4%

Please cite this article in press as: Cronin, M. A. et al., Why don't well-educated adults understand accumulation? A challenge ..., *Organizational Behavior and Human Decision Processes* (2008), doi:10.1016/j.obhdp.2008.03.003

A. Priming Task



Write down how many people are in the store each minute:

Minute 1:
Minute 2:
Minute 3:
Minute 4:
Minute 5:

B. Simple SF Task (Participants Were Asked the Same Four Questions S

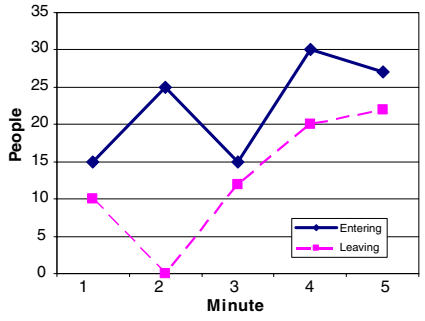


Fig. 4. Experiment 4.

task, participants were given the simple task shown in Fig. 4B and asked to answer the standard four questions.

Results

As in prior experiments, participants did well on the first two questions (Table 5), indicating that they could read the graph, but generally did not understand the concepts of accumulation (27% and 38% for Questions 3 and 4, respectively). Success on the stock–flow questions was marginally higher than that of participants who received the task without the priming condition, the baseline for this population (8% and 16% on an isomorphic 5-point graph),[6] so priming did have some effect, partially supporting Hypothesis 5. Yet priming with did not eliminate the problem. Nearly half (18 of 37) did the priming task incorrectly despite its simplicity; most of these participants responded that the number of people in the store each minute was 5, 5, 5, 5, 5—that is, they gave the net flow of people into the store each minute rather than the total number (5, 10, 15, 20, 25). Those who responded correctly on the priming task did significantly better on the stock–flow questions in Fig. 4B than those who did not: None of those who failed on the priming task correctly identified when the most people were in the store, compared to about half of those who did the priming task correctly ($p = .0004$). Only one of those who failed on the priming task correctly identified when the fewest people were in the store, compared to 68% of those who got the priming task right ($p = .0001$).

The results suggest that many (about half the participants) did not understand the concept of accumulation. However, even for those who answered the priming question correctly, success rates on the stock–flow questions in Fig. 4B remained discouragingly low. Many who could accumulate the net flow of people in the store in the priming task (Fig. 4A) were unable to determine when the most and fewest people were in the store in Fig. 4B, despite the extreme simplicity of that task. It appears that many participants not only had difficulty applying the principles of accumulation but also failed to recognize the stock–flow structure, even after being explicitly directed to carry out the accumulation of inflow and outflow into a stock. The results suggest that, for these people at least, the problem is not the failure to activate latent knowledge of accumulation but the lack of such knowledge.

Experiment 5: The correlation heuristic

The experiments reported here verify that SF failure is a robust error. Many people do not understand the principles of accumulation, or fail to apply their knowledge, despite explicit instructions. What heuristics, then, do they use? Prior work (Booth Sweeney & Sterman, 2000) suggests that many people use a "correlation heuristic", a form of pattern matching in which people assume that the output of a system (e.g., the level of water in a tub) should "look like" the input (the flow or net flow of water into the tub). Booth Sweeney and Sterman (2000) found extensive use of the correlation heuristic among erroneous responses to simple tasks such as inferring the level of water in a tub from graphs of the flow in and drain out or inferring the cash balance of a firm from graphs of receipts and expenditures. These results have been replicated with diverse student populations (e.g., Atkins et al., 2002; Ossimitz, 2002; Pala & Vennix, 2005). However, these studies raise a number of questions. If people correlate inputs to outputs, what cues do they select—the inflow, outflow, or net flow—and why? Do people focus on trends or on specific points such as the maximum or minimum of the flows? How does the information display affect the use of the correlation heuristic?

To explore these questions, we first coded subject responses in Experiments 1, 2 and 3 for evidence of correlational reasoning.[7] In these experiments, the correlation heuristic suggests that the maximum of the stock coincides with the maximum of the inflow (or net inflow), and the minimum of the stock with the maximum of the outflow (or net outflow). Table 6 shows the frequency of particular erroneous responses to the stock–flow questions (when are the most/fewest people in the store?), including: the first point on the graph (*Start*); the final point (*End*); the maximum inflow (*Peak Inflow*); the maximum outflow (*Peak Outflow*); the maximum net inflow (*Peak Net Inflow*); the maximum net outflow (*Peak Net Outflow*); the point where inflow and outflow cross (*Cross*), "can't be determined" (*CBD*); the minimum of the inflow (*Valley*); and all others (*Other*). The experiments were carefully designed to distinguish among these options. For example, in the baseline task (Fig. 1, Table 1), the peak inflow occurs at $t = 4$, peak net inflow at $t = 8$, the crossing point at $t = 13$, peak net outflow at $t = 17$ and peak outflow at $t = 21$.

The most frequent erroneous response for the maximum stock question is the maximum net inflow, and the most frequent erroneous response for the minimum stock question is the maximum net outflow (Table 6). The pattern holds across a range of cover stories and protocols and across the participants from MIT, Carnegie

[6] Though the baseline graph was given in a different semester, the populations from which all the George Mason University participants were drawn were very similar, as were recruitment methods and the manner (room, time, etc.) in which the tasks were administered. Nevertheless, the same caveat as in note 4 applies.

[7] Experiment 4 used a graph showing data for only 5 min, so the different critical points cannot be distinguished.

Please cite this article in press as: Cronin, M. A. et al., Why don't well-educated adults understand accumulation? A challenge ..., *Organizational Behavior and Human Decision Processes* (2008), doi:10.1016/j.obhdp.2008.03.003

Table 5
Experiment 4: Influence of priming on success

	Question 1: Most entering?	Question 2: Most leaving?	Question 3: Most in store?	Question 4: Fewest in store?
Baseline ($N = 37$)	95%	92%	8%	16%
Priming condition ($N = 37$)	86%	89%	27%	38%
Exact test (p)	.430	.999	.063	.065
Prime correct ($N = 19$)	95%	95%	53%	68%
Prime incorrect ($N = 18$)	78%	83%	0%	6%
Exact test (p)	.180	.340	.0004	.0001

Mellon, and George Mason. Few selected the maximum of the gross inflow (outflow) as the point where the most (fewest) are in the store, indicating that most recognized the importance of the net rate of change in determining the stock. Yet they nevertheless concluded that the maximum (minimum) of the stock coincides with the maximum (minimum) of the net flow, consistent with the correlation heuristic.

We hypothesize that the tendency to correlate the maximum of the net inflow (outflow) to the maximum (minimum) of the stock depends on the prominence of these points in the presentation of the data. What determines the salience of the maximum net inflow/net outflow? These points are readily determined in the line graph format, harder to see in the bar graph format, and still more difficult to discern in the tabular and text formats where one must calculate the net change each minute and then compare them (compare Figs. 2A–D). We therefore expect:

$H_{5.1}$: The fraction of participants erroneously selecting the maximum net inflow (outflow) for the maximum (minimum) of the stock, respectively, will be greatest in the line graph format and lowest in the tabular and text-based formats.

Table 6
Frequency of incorrect responses on the stock–flow questions

	Maximum value of stock? (%)			Minimum value of stock? (%)		
	Exp. 1 MIT	Exp. 2 CMU	Exp. 3 GMU	Exp. 1 MIT	Exp. 2 CMU	Exp. 3 GMU
Start point	0	0	0	10	3	0
End point	1	8	0	*	*	*
Peak inflow	6	3	15	1	0	4
Peak outflow	1	0	4	3	0	4
Peak net inflow	52	33	54	0	3	0
Peak net outflow	1	18	0	43	18	35
Crossing point	*	*	*	3	18	8
Valley	0	0	0	0	0	15
Cannot be determined	30	40	23	36	53	31
Other	10	0	4	4	0	4

* Correct response; table reports frequencies of incorrect responses.

Table 7 shows the results for Experiment 1 (Fig. 2). The proportion of those answering incorrectly who assert that the maximum (minimum) of the stock occurs at the maximum net inflow (outflow) is significantly higher for the line graph compared to the bar graph (exact test $p = .016$). The proportion of participants answering incorrectly who assert that the maximum (minimum) of the stock occurs at the maximum net inflow (outflow) is also significantly higher in the two graphical conditions (line and bar) compared to the two non-graphical (table and text) conditions (exact test $p = .0003$). $H_{5.1}$ is supported, indicating that the use of correlational reasoning increases with the salience of the maxima in the net flow data.

The analysis above strongly suggests that participants with weak understanding of the principles of accumulation tend towards use of the correlation heuristic. However, in all the experiments reported thus far, participants select a value for the time at which the stock reaches its extreme values; these responses do not allow us to determine whether people believe that the trajectory of the stock matches that of the flows throughout the time horizon. The final experiment directly assesses the prevalence of correlational reasoning by presenting people with graphs of the inflow and outflow to a stock and asking them to draw the trajectory of the stock.

Method

Participants ($N = 282$) were students enrolled in a subsequent term of the same course at the MIT Sloan School of Management used in the baseline experiment and Experiment 1. The participants were demographically similar to the prior groups: average age was 28 (range 20–44), and 71% were male; 54% were trained in STEM; 37% were trained in economics or other social science; and 29% held prior advanced degrees. The protocol was identical to that of Experiment 1, except that participants were randomly assigned to one of eight treatment conditions (Fig. 5). Each shows a graph displaying the flow of people entering and leaving a store over 30 min. Participants were directed to draw the number of people in the store throughout the 30 min on a blank graph placed directly beneath the flow graph. The eight flow patterns ranged from the exceptionally simple (constant flows) to more complex shapes. Note that no numerical scale is provided for the flow data, or for the blank graph

Table 7
Frequency of incorrect responses in Experiment 1

	Maximum value of stock? (%)				Minimum value of stock? (%)			
	Line graph	Bar graph	Table	Text	Line graph	Bar graph	Table	Text
Start	0	5	13	0	18	27	16	14
End	0	5	0	12	*	*	*	*
Max inflow	4	0	6	12	0	0	0	0
Max outflow	0	0	3	0	0	7	3	0
Max net inflow	44	16	3	6	0	3	0	4
Max net outflow	4	0	0	9	29	13	0	18
Crossing point	*	*	*	*	3	0	0	4
CBD	48	68	58	48	47	47	58	57
Other	0	5	16	12	3	3	23	4

* Correct response; table reports frequencies of incorrect responses.

Please cite this article in press as: Cronin, M. A. et al., Why don't well-educated adults understand accumulation? A challenge ..., *Organizational Behavior and Human Decision Processes* (2008), doi:10.1016/j.obhdp.2008.03.003

Participants received one of eight patterns for the inflow and outflow to the stock. The example below shows treatment 4; Figure 6 shows all eight patterns.

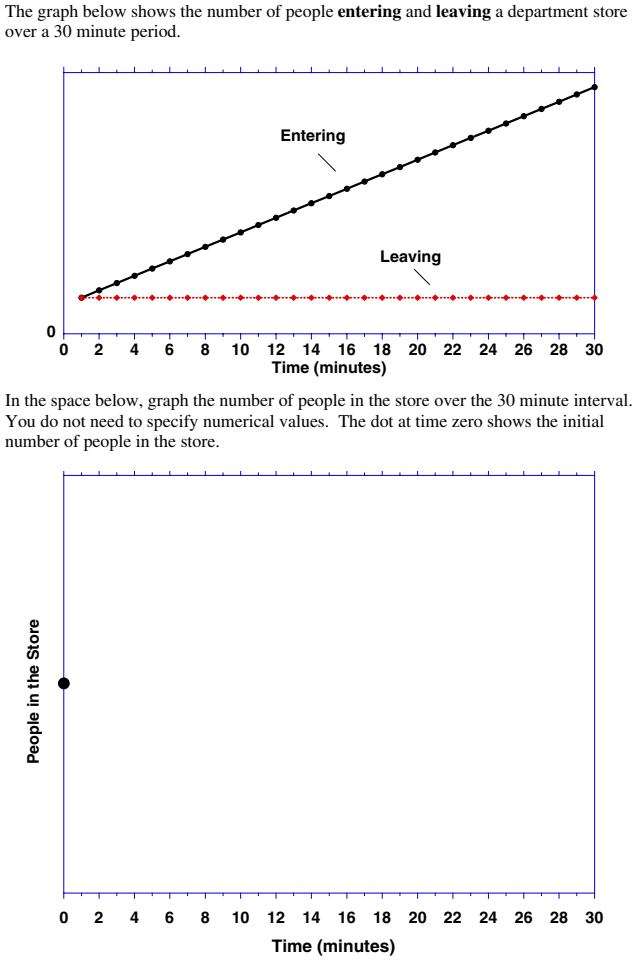


Fig. 5. Experiment 5. Testing the correlation heuristic.

for their response. The graph for the stock includes a point indicating the initial number of people in the store. To avoid biasing participant responses, that point is placed at the midpoint of the vertical axis. In all cases, it is possible to answer correctly without knowledge of calculus and without carrying out any calculations.

To code the responses, we first determined whether participants sketched a pattern that was qualitatively correct. A pattern was judged qualitatively correct if it was consistent with basic stock–flow principles: (i) the stock is rising, constant, or falling when the net inflow is positive, zero, or negative, respectively; and (ii) the rate of change (slope) of the stock is increasing (decreasing) when the net flow is increasing (decreasing). Participants were not penalized for drawing patterns that were not quantitatively correct or that did not show the number in the store beginning at the initial point provided on the graph. We then coded the erroneous responses to determine whether the path they drew matched the pattern of the inflow, outflow, or net flow, that is, whether the correlation between the stock and inflow or net flow was +1 (perfect pattern matching), or −1. A correlation of −1 indicates perfect pattern matching, but with the pattern inverted; such inversion might occur when the net flow is positive but falling (e.g., treatment 5); in such a case the participant realizes that the stock is rising, but still erroneously concludes the stock follows the shape of the net flow.

The eight flow patterns divide into three groups. Group I consists of treatments 1 and 2 and should be the easiest: participants need only realize that the net flow is constant, determine whether it is positive or negative, and draw a straight line with positive or negative slope. Group II consists of treatments 3, 4 and 5. These all have constant outflow and linear inflow: participants must determine whether the net flow is positive or negative, note whether the net flow is increasing or decreasing, and then draw a curve that is rising or falling at an increasing or decreasing rate. Group III comprises treatments 6, 7, and 8 and should present the greatest difficulty: These have constant outflows but nonlinear patterns for the inflow: participants must determine whether the net flow is positive or negative, then determine whether the net flow is increasing or decreasing in each part of the 30-min interval, and sketch a path that shows the stock rising or falling with qualitatively correct changes in slope. Hence:

$H_{5.2}$: Performance (Group I) > Performance (Group II) > Performance(Group III).

The more difficult the task, the greater the likelihood people will revert to the use of a heuristic rather than reasoning through the task. Here, difficulty increases with the complexity of the patterns for the flows. We therefore expect that the use of the corre-

Please cite this article in press as: Cronin, M. A. et al., Why don't well-educated adults understand accumulation? A challenge ..., *Organizational Behavior and Human Decision Processes* (2008), doi:10.1016/j.obhdp.2008.03.003

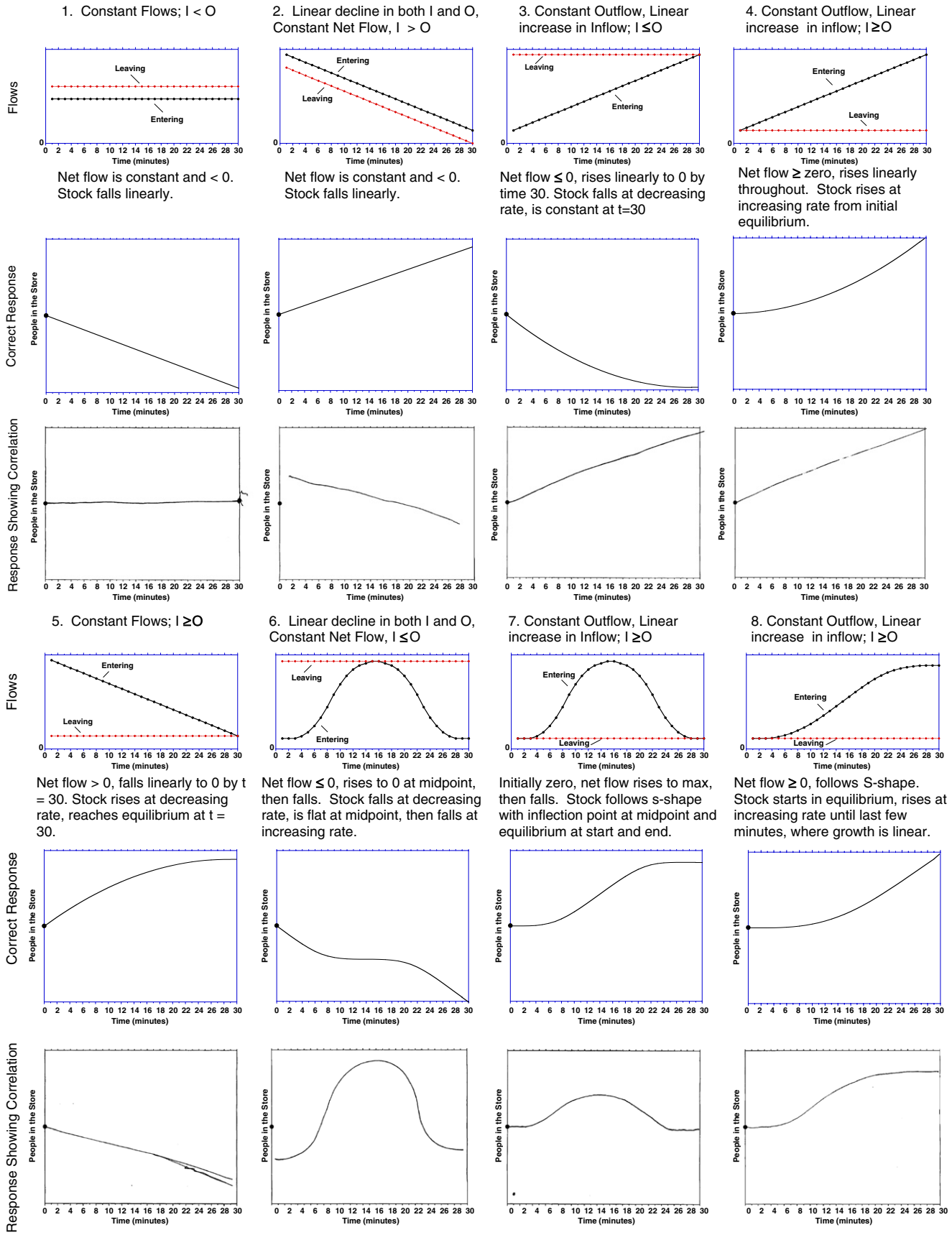


Fig. 6. Correct and typical incorrect responses for Experiment 5.

lation heuristic will increase with the difficulty of the tasks. Letting F_{corr} indicate the fraction of erroneous responses in which participants draw patterns exactly matching the shape of the inflow or net inflow,

$H_{5.3}$: F_{corr} (Group I) < F_{corr} (Group II) < F_{corr} (Group III).

Results

Despite the simplicity of the flow patterns, performance was poor (Table 8). Overall, only 54% drew the correct pattern. The proportion correct varies substantially across the treatments, from 83% correct in treatment 1, with constant flows, to 19% correct in treatment 8, where the outflow is constant and the inflow follows an S-shape with $I \geqslant O$ at all times.

Most interesting, the majority of those responding incorrectly matched the pattern of the inflow or net flow. Overall, 71% of the incorrect responses show paths for the stock identical in shape to that of the inflow or net flow (that is, the correlation between the stock and inflow or net flow = +1). Fig. 6 shows the correct paths for the stock and a typical incorrect response for each of the treatments. As hypothesized, treatment 1, where the flows are constant, yields the highest overall performance (83% correct) and lowest incidence of correlation among incorrect responses (33%). The highest incidence of correlation (89% of incorrect responses) arises in treatments 4 and 8. In both of these the stock rises at an increasing rate as the net flow grows over time, but 89% of those responding incorrectly drew patterns that matched the net flow (or, equivalently, the inflow). $H_{5.2}$ is supported: performance on Group I is significantly better than that for Group II ($p = .0005$), and performance on Group II is significantly better than that for Group III ($p = .012$). Hypothesis 5.3 is partially supported: In Group I, 47% of erroneous responses are perfectly correlated with the inflow or net flow, significantly lower than the 80% of erroneous responses in Group II exhibiting correlation ($p = .021$). However, the fraction of responses showing correlation in Groups II and III do not differ significantly.

The individual treatments within the groups provide additional insight into the conditions leading people to use the correlation heuristic. Performance on treatment 1 is not significantly better than that for treatment 2. Few in this participant pool, with their strong technical backgrounds, found these simple patterns, where the net flow is constant, to be difficult. Most of those who erred, however, drew a horizontal line, indicating failure to accumulate the net change into the stock.

Table 8
Results of Experiment 5

Treatment	% Incorrect	% Incorrect exhibiting correlation	Corr(stock, inflow) = +1 (%)	Corr(stock, net flow) = +1 (%)	*N*
1[a]	16.7	33.3	33.3	33.3	36
2[b]	22.2	55.6	11.1	44.4	37
3	41.7	68.8	43.8	25.0	37
4	55.6	88.9	88.9	88.9	34
5	44.4	80.0	33.3 Corr(S, I) = −1: 46.7	33.3 Corr(S, N) = −1: 46.7	35
6	69.4	56.0	24.0	32.0	36
7	47.2	57.1	57.1	57.1	33
8	80.6	88.9	88.9	88.9	34
All	46.1	70.8	53.1	54.6	282

[a] Inflow, outflow and net flow are all constant. "Corr(stock, inflow) = +1" indicates that the response was also constant (a horizontal line).

[b] Inflow and outflow are correlated; net flow is constant. "Corr(stock, net flow) = +1" indicates that the response was also constant (a horizontal line).

In treatment 3, where the inflow rises but the stock falls because Outflow > Inflow, 42% drew incorrect patterns, with 44% of these incorrect responses matching the pattern of the inflow and 25% matching the pattern of the net flow (showing a straight line with negative slope). In treatment 4, 56% gave incorrect responses, nearly all (89%) drawing a positively sloped straight line that matched the pattern of the inflow (and net flow), while the stock actually rises at an increasing rate. In treatment 5, where the inflow falls but the net flow is positive (so the stock rises), 33% of incorrect responses were straight lines with negative slope, i.e., were perfectly correlated with the inflow (or net flow). However, 47% of incorrect responses were straight lines with positive slope. These participants realized that the stock was increasing, but then were not able to apply the principles of accumulation and instead relied on correlational reasoning.

Participants receiving treatment 6, where the net flow is negative, outperformed those receiving treatment 7, where the net flow is positive (31% vs. 53% correct; $p = .03$). The fraction of incorrect responses matching the pattern of the inflow or net flow in these conditions is nearly equal. Many people incorrectly follow the pattern of the inflow or net flow rather than using information on the sign of the net flow to determine whether the stock is rising or falling. These erroneous judgments lead to patterns implying that the number of people in the store can rise even though more people are leaving than entering, or, equivalently, that the level of water in a tub can rise even when water drains out faster than it flows in.

General discussion

Results from the experiments reported here demonstrate an important and pervasive problem in human reasoning: our inability to understand stocks and flows, that is, the process by which the flows into and out of a stock accumulate over time. Stock and flow structures are pervasive in systems at all scales, from the accumulation of water in a tub to the accumulation of greenhouse gases in the atmosphere. Effective decision making in dynamic settings requires decision makers to understand accumulation. Prior work has demonstrated that even highly educated people do poorly on a range of simple stock–flow problems. This research demonstrates that SF failure is not an artifact of the task, nor is it easily corrected. Rather, the error reflects serious misunderstanding of the basic principles of accumulation.

We tested whether people in fact understand the concepts of accumulation, but perform poorly due to information displays, unfamiliar contexts, inadequate motivation, inability to read or construct graphs, or limited cognitive capacity. Poor performance persisted among highly educated individuals with strong training in science, technology, engineering and mathematics even when the tasks could be done without any calculation, when the number of data points presented was reduced by 60% and regardless of whether the data are displayed in line graphs, bar graphs, tables, or text (Experiment 1). Poor performance was robust to changes in the cover story and to contexts that involved discrete entities or continuously varying quantities (Experiment 2). Modest incentives to respond correctly did not lead to improvement (Experiment 3). Many could not correctly accumulate the quantity in the stock even when they were explicitly directed to do so in a problem with constant flows (Experiment 4). Although outcome feedback indicating when participants had provided an incorrect answer did improve performance, the improvement was slow, and a number of people never responded correctly, even after many trials (Experiment 3). Finally, nearly half of a sample of highly educated graduate students with extensive technical training were unable to correctly draw the qualitative path of a stock from very simple patterns for its inflow and outflow, and roughly

Please cite this article in press as: Cronin, M. A. et al., Why don't well-educated adults understand accumulation? A challenge ..., *Organizational Behavior and Human Decision Processes* (2008), doi:10.1016/j.obhdp.2008.03.003

70% instead drew patterns that matched those of the inflow or net flow (Experiment 5).

Although most of the experiments allowed participants 10 min to finish the task, most of the participants finished much earlier. Many reported high confidence that their answers were correct, even when they were not. For example, in Experiment 3, where participants received performance feedback, many expressed disbelief when told that their answers were incorrect. These behaviors, coupled with the persistence of poor performance in the face of large manipulations in task features, context, and so forth, suggest that SF failure shares some features with insight problems (Mayer, 1995). Insight problems are analytically easy—once one recognizes the proper frame to use. Until then, people tend to use a flawed but intuitively appealing (and hence difficult to change) problem frame.

People appear to employ heuristics that are intuitively appealing but erroneous—specifically, many use the correlation heuristic, reasoning that the output of the system (here, the stock) should "look like" (be highly correlated with) its inputs (here, the flows or net inflow). Such individuals fail to grasp the fundamental principle that any stock rises (falls) when the inflow exceeds (is less than) the outflow. The intuitive appeal of the correlation heuristic appears to be quite strong: attempts to activate whatever latent knowledge of accumulation participants may have through cover stories emphasizing familiar contexts with continuously varying quantities (Experiment 2), through motivation and feedback (Experiment 3), and by explicitly directing people to accumulate a stock prior to doing the task (Experiment 4) had little impact. Most of the erroneous responses are consistent with the use of the correlation heuristic. Over a range of experiments and participant populations, a plurality of those asked to identify when a stock reaches its maximum (minimum) select the point with the highest net inflow (net outflow), consistent with correlational thinking but violating fundamental principles of accumulation. Further, the frequency of use of the correlation heuristic increases as the flows become more complex than simple straight lines (Experiment 5).

Future work should investigate the cues that trigger or inhibit the use of the correlation heuristic and the learning processes through which individuals acquire and use the deep structure of the problem. Verbal protocols, as Chi, Feltovitch, & Glaser, 1981 suggest, may help reveal the nature and construction of people's mental representations as they try to discover the relationships between stocks and flows. Other methods may be needed to the extent learning in dynamic tasks is implicit so that individuals are unable to verbalize the ways in which they make decisions (Gonzalez, 2005b; Gonzalez, Lerch, & Lebiere, 2003). Interactive decision-making environments may be particularly useful in the investigation of the SF failure, both to learn how people make decisions and to speed learning (Gonzalez & Dutt, 2008; Sterman 2002). Rather than making judgments on a static task, using interactive decision-making environments, people continuously or periodically make decisions and receive feedback as they observe how the state of the system responds (Diehl & Sterman 1995; Sterman 1989a; 1989b). For example, individuals could make decisions about the flows affecting a stock each minute, observe the resulting value of the stock, then make flow decisions for the next minute, and so on (Gonzalez & Dutt, 2008).

Research should also explore the extent to which SF failure might be reinforced by the educational system. Investigating the role of formal schooling in the genesis of SF failure is important for organizational behavior and decision making among adults: the greater the extent to which early education inadvertently reinforces SF failure the harder it will be to overcome by the time people begin to make consequential decisions in systems involving accumulations. Formal education could reinforce SF failure in two ways: by encouraging the use of correlational reasoning and by failing to teach the principles of accumulation. Educators have documented errors in mathematics problem solving involving the erroneous use of correlational reasoning (Ben-Zeev & Star, 2001; Harel, Behr, Post, & Lesh, 1992; Van Dooren, De Bock, Hessels, Janssens, & Verschaffel, 2005). Early mathematics education provides students with extensive practice in proportional reasoning and they are often encouraged to think linearly (Van Dooren et al., 2005). For example, children often encounter problems emphasizing proportionality such as 'if 2 cups of water fill one bucket, how many cups fill three buckets?' (van den Brink & Streefland, 1979). Such training may reinforce the impression that relations between variables are proportional (Van Dooren, De Bock, Janssens, & Verschaffel, 2007). Mathematics education may strengthen any predisposition to use the correlation heuristic people have prior to formal schooling. Further research should help disentangle interactions between innate cognitive structures, schooling, and other experiences in the genesis of SF failure and related difficulties in dynamic decision making.

The greatest challenge for future work is to find effective methods to improve performance on SF problems, improving our ability to understand and manage the complex systems affecting our personal lives, organizations, and society.

Acknowledgments

This research was partially supported by the National Science Foundation (Human and Social Dynamics: Decision, Risk, and Uncertainty, Award No. 0624228) and by the Army Research Laboratory (DAAD19-01-2-0009) awards to Cleotilde Gonzalez. John Sterman is co-PI of and supported by the Project on Innovation in Markets and Organizations at the MIT Sloan School of Management. We are grateful to Polina Vanyukov, who helped collect some of the data reported here. We also thank Jeff Loewenstein, Don Moore, Cathy Tinsley, and anonymous reviewers for their comments.

References

Almor, A., & Sloman, S. A. (2000). Reasoning versus memory in the Wason selection task—A non-deontic perspective on perspective effects. *Memory and Cognition, 28*, 1060–1069.

Atkins, P., Wood, R., & Rutgers, P. (2002). The effects of feedback format on dynamic decision making. *Organizational Behavior and Human Decision Processes, 88*, 587–604.

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

Berry, J., & Nyman, M. (2003). Promoting students' graphical understanding of the calculus. *Journal of Mathematical Behavior, 22*, 481–497.

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

Brehmer, B. (1990). Strategies in real-time, dynamic decision making. In R. M. Hogarth (Ed.), *Insights in decision making* (pp. 262–279). Chicago: University of Chicago Press.

Brehmer, B. (1995). Feedback delays in complex dynamic decision tasks. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 103–130). Hillsdale, NJ: Lawrence Erlbaum Associates.

Carlson, M., Jacobs, S., Coe, E., Larsen, S., & Hsu, E. (2002). Applying covariational reasoning while modeling dynamic events: A framework and a study. *Journal for Research in Mathematics Education, 33*(5), 352–367.

Camerer, C., & Hogarth, R. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty, 19*(1–3), 7–42.

Chi, M. T. H., Feltovitch, P. J., & 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.

Dawes, R. M. (1988). *Rational choice in an uncertain world*. Fort Worth, TX: Harcourt Brace.

Dawes, R. M. (1998). Behavioral decision making and judgment. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., pp. 497–548). New York: McGraw-Hill.

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

Gattis, M. (2002). Structure mapping in spatial reasoning. *Cognitive Development, 17*, 1157–1183.

Gattis, M. (2004). Mapping relational structure in spatial reasoning. *Cognitive Science, 28*, 589–610.

Gattis, M., & Holyoak, K. J. (1996). Mapping conceptual to spatial relations in visual reasoning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 22*(1), 231–239.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition, 43*, 127–171.

Gonzalez, C. (2005a). Decision support for real-time dynamic decision making tasks. *Organizational Behavior & Human Decision Processes, 96*, 142–154.

Gonzalez, C. (2005b). The relationship between task workload and cognitive abilities in dynamic decision making. *Human Factors, 47*(1), 92–101.

Gonzalez, C., & Dutt, V. (2008). On controlling the simplest dynamic system, in preparation.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science, 27*(4), 591–635.

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

Houghton, J. T., Ding, Y., Griggs, D. J., Noguer, M., van den Linden, P. J., & Dai, X., et al. (Eds.). (2001). *Climate change 2001: The scientific basis. Contribution of working group I to the third assessment report of the Intergovernmental Panel on Climate Change*. New York: Cambridge Unviersity Press.

IPCC (2007). *Climate change 2007: The physical science basis*. Geneva, Switzerland: Intergovernmental Panel on Climate Change. Available from www.ipcc.ch.

Jensen, A., & Brehmer, B. (2003). Understanding and control of a simple dynamic system. *System Dynamics Review, 19*(2), 119–137.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430–454.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*, 211–228.

Kleinmuntz, D. N. (1985). Cognitive heuristics and feedback in a dynamic decision environment. *Management Science, 31*(6), 680–702.

Kleinmuntz, D. N., & Schkade, D. A. (1993). Information displays and decision processes. *Psychological Science, 4*(4), 221–227.

Mayer, R. E. (1995). The search for insight: Grappling with Gestalt psychology's unanswered questions. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 3–32). Cambridge, MA: MIT Press.

Omodei, M., & Wearing, A. (1995). The Fire Chief microworld generating program: An illustration of computer-simulated microworlds as an experimental paradigm for studying complex decision-making behavior. *Behavior Research Methods, Instruments, & Computers, 27*, 303–316.

Ossimitz, G. (2002). Stock–flow-thinking and reading stock–flow-related graphs: An emperical investigation in dynamic thinking abilities. Paper presented at the International System Dynamics Conference.

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

Paulos, J. (1988). *Innumeracy: Mathematical illiteracy and its consequences*. New York: Hill and Wang.

Petty, R. E., & Wegener, D. T. (1998). Matching versus mismatching attitude functions: Implications for scrutiny of persuasive messages. *Personality and Social Psychology Bulletin, 24*(3), 227–240.

Petty, R. E., Wegener, D. T., & Fabrigar, L. R. (1997). Attitudes and attitude change. *Annual Review of Psychology, 48*, 609–647.

Raman, M. (2002). Coordinating informal and formal aspects of mathematics: Student behavior and textbook messages. *Journal of Mathematical Behavior, 21*, 135–150.

Roch, S. G., Lane, J. A. S., & Samuelson, C. D. (2000). Cognitive load and the equality heuristic: A two-stage model of resource overconsumption in small groups. *Organizational Behavior and Human Decision Processes, 82*, 185–212.

Simon, H. A. (1979). *Models of thought*. New Haven, CT: Yale University Press.

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

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

Sterman, J. D. (2002). All models are wrong: Reflections on becoming a systems scientist. *System Dynamics Review, 18*, 501–531.

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

Sterman, J. D., & Booth Sweeney, L. (2007). Understanding public complacency about climate change: Adults' mental models of climate change violate conservation of matter. *Climatic Change, 80*(3–4), 213–238.

Tufte, E. (1983). *The visual display of quantitative information*. Cheshire, Connecticut: Graphics Press.

Tufte, E. (1990). *Envisioning information*. Cheshire, Connecticut: Graphics Press.

Van den Brink, J., & Streefland, L. (1979). Young children (6–8)—Ratio and proportion. *Educational Studies in Mathematics, 10*, 403–420.

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.

Van Dooren, W., De Bock, D., Janssens, D., & Verschaffel, L. (2007). Pupils' over-reliance on linearity: A scholastic effect? *British Journal of Educational Psychology, 77*, 307–321.

Vicente, K. J. (1996). Improving dynamic decision making in complex systems through ecological interface design: A research overview. *System Dynamics Review, 12*(4), 251–279.

Vicente, K. J., & Rasmussen, J. (1992). Ecological interface design: Theoretical foundations. *IEEE Transactions of Systems, Man, and Cybernetics, 22*(4), 589–606.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*, 129–140.

Wickens, C. D., & Carswell, C. M. (1995). The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors, 37*(3), 473–494.

Please cite this article in press as: Cronin, M. A. et al., Why don't well-educated adults understand accumulation? A challenge ..., *Organizational Behavior and Human Decision Processes* (2008), doi:10.1016/j.obhdp.2008.03.003