Does a Better Understanding of Accumulation Indeed Predict a Higher Performance in Stock and Flow Management?

Preliminary Results Based upon a Pilot Study

Jürgen Strohhecker

Frankfurt School of Finance and Management

Sonnemannstraße 9-11, 60314 Frankfurt am Main

Telephone: +49 69 154008-110

E-Mail: j.strohhecker@frankfurt-school.de

ABSTRACT

The stock and flow management (SFM) problem is of high relevance for a broad range of decision makers in society, business, and personal affairs. Although in some areas highly sophisticated models and control concepts have been developed, the phenomena of excess stock and shortages are omnipresent. One recent explanation for these observations is offered by a stream of research, which finds evidence for widespread and persistent deficits in stock-flow thinking (SFT) capabilities even among well-educated adults. Building on this explanation, an attempt is made to test the hypothesis that the better people understand accumulation, the higher will be their performance in SFM tasks. However, the results of a small sample pilot study indicate falsification. Therefore, Ackerman's PPIK theory is introduced and used to formulate an elaborate causal model, which could be tested in future research.

Key Words: Accumulation, Bathtub Dynamics, Dynamic decision making, Intelligence, Stocks and flows, Stock-flow failure, System Dynamics

INTRODUCTION – THE STOCK AND FLOW MANAGEMENT PROBLEM

Stock and flow management is of high relevance for a broad range of decision makers in society, business, and personal affairs: A Secretary of Finance, for example, has to manage tax inflows and public expenditure outflows which accumulate in the national debt or asset stock. Purchasing managers have to decide when to order which quantities so that both stock-outs and inventory costs are minimized. In terms of an everyday application, each human being is obliged to manage both the intake and burning of calories to maintain a healthy weight. Anecdotal evidence of poor SFM performance is omnipresent in everyday life. TV series increasingly portray unlucky fellow human beings who have accumulated such a high amount of debt by overusing credit cards or consumer credits that they will never ever be able to pay off their pile of debt. Finally, the financial crisis of 2008 and 2009 shows that even bank managers struggle to keep their banks' liquidity in balance when capital markets are disrupted and the economic environment radically changes.

Inventory management is another well investigated field with ample examples of poor SFM performance. For example, it happens regularly that one goes to a supermarket in order to replenish the refrigerator at home only to find the milk shelf empty. As surveys show, out-of-stock products rank indeed high in the top list of customer annoyances (2008a; 2008b). Severe stock-out or parts shortage phenomena are even reported in the news – for example, Bosch's delivery problems of diesel-fuel injection pumps, which caused production stops at Mercedes, BMW, Audi, and Opel (Milne, 2005), or Apple's iPhone shortage (Hansell, 2008). Such shortages can severely damage earnings and raise costs.

Excess inventories are, however, just as bad as stock-outs. Large inventories increase the working capital and the inventory risk. When goods perish before they can be sold at the regular price – either literally in the sense of food spoiling or figuratively in the sense that products could go out of fashion – shareholder value is obviously destroyed (e.g., Foster, 2004). In the worst-case-scenario of extreme excess inventory coinciding with decreasing demand, a company's financial solvency is endangered and bankruptcy will be the consequence.

For complex systems, dynamic decision making research has accumulated evidence of SFM failures by conducting a broad range of decision making experiments. (e.g., Ackerman & Kanfer, 1993; Brehmer, 1995; Diehl & Sterman, 1995; Kleinmuntz, 1985; Reichert & Dörner, 1988). This work suggests that human beings have severe difficulties understanding and managing systems which are dynamically complex, that is, which are characterized by feedback, time delays, nonlinearities, and accumulation. Sterman's (1989a; 1989b) pioneering behavioural supply chain research that made use of the Beer Game (Senge, 1990) as an experimental setting shows that the subjects' inventory management performance suffers systematically from the use of inappropriate anchoring heuristics and misperceptions of time lags. Croson and Donohue (2003; 2006) build on this research by confirming that low – albeit improved – supply chain performance still exists when participants are aware of the underlying demand distribution or point of sales data. Bloomfield, Gino, and Kulp (2007) find lamentable results even in a single echelon supply chain experiment, where inter-echelon coordination problems are absent. In summary, prior work suggests that dynamic decision making tasks, regardless whether complex or "simple," represent real challenges for human beings.

Moreover, recent work has revealed that a large fraction of highly educated people is unable to infer the behaviour of even the most simple stock-flow-systems consisting of only one stock, one inflow, and one outflow (Booth Sweeney & Sterman, 2000). As no feedback, no time delays, or nonlinearities were incorporated in those simplistic systems, they cannot be characterized as dynamically complex. Nevertheless, the average understanding of these systems' dynamic is lamentable. The subjects showed a rather poor performance in a variety of paper-and-pencil tasks involving such systems, which supports the conclusion that human beings indeed have a poor understanding of accumulation. Subsequent studies by Ossimitz (2002), Sterman and Booth Sweeney (2002; 2007), Cronin and Gonzales (2007) corroborate the hypothesis that poor SFT is a persistent phenomenon, comparable to the deep-rooted problems people have in

probabilistic judgements and decision making (Hastie & Dawes, 2001; Kahneman & Tversky, 1972).

This research attempts to contribute to both research streams outlined above by linking them with the hypothesis that SFT capabilities causally affect SFM performance. This hypothesis is put to the test by making two observations in a laboratory setting. The first observation setting uses a collection of three stock and flow tasks that were used in prior studies for assessing SFT ability. The second setting employs a dynamic, yet simple inventory management game with stochastic demand, a four period lead time, and costs for ordering, inventory keeping, and stock outs. The performance is assessed by subtracting the cumulated costs from two benchmarks derived from applying two optimized order-quantity (Q, r) rules. A pilot study is conducted to test the research design. The results of this pilot indicate that the hypothesis in its simple, one-dimensional form has to be rejected and replaced by a more elaborate cause-and-effect model.

The paper continues in Section 2 with a description of the hypothesis to be tested and the research method used. Section 3 describes how the SFT ability is measured and outlines the results. Section 4 provides details on the inventory management task and the assessment of the subjects' performance. Section 5 presents the results of the hypothesis test, and Section 6 introduces an advanced causal model derived from Ackerman's PPIK theory. The paper concludes with a discussion of limitations and contributions of this research and outlines directions for further research.

HYPOTHESIS AND RESEARCH METHOD

Prior work has revealed that people perform rather badly in both rather complex (Croson et al., 2003, 2006; Sterman, 1989a, b) and rather simple SFM tasks (Bloomfield et al., 2007). In searching for the simplest dynamic task that people can cope with, Sterman and others developed paper-and-pencil tasks based upon the simplest system possible with one inflow, one stock, and one outflow, with no feedback, time delays, and non-linearity, and found that even well educated subjects still struggle with the understanding of stocks and flows (Booth Sweeney et al., 2000; Cronin, Gonzalez, & Sterman, 2009; Ossimitz, 2002; Sterman, 2002). Cronin et al. (2009) find that poor SFT performance persists regardless of whether the data are displayed in line graphs, bar graphs, tables, or text; poor performance is robust to changes in the cover story that frames the task and provides a specific context, for example the management of a stock of cash or the amount of water in a bathtub; it is also robust to situations that involve discrete entities or continuously varying quantities; even reducing the task complexity by decreasing the number of data points presented does not increase SFT performance. SFT capabilities obviously suffer from important and pervasive shortcomings in human reasoning. A high percentage of people seriously misunderstands "the basic principles of accumulation" (Cronin et al., 2009).

Whereas the reasoning of Sterman (2002), Cronin et al. (2009) and others is plausible, that is, that poor SFM performance is related to poor stock flow understanding, as far as I know, this hypothesis has not yet been put to the test. The objective of this research is to contribute to the literature by formulating and testing the following hypothesis:

Hypothesis 1. The better people understand stocks and flows, the better they perform in managing a stock and flow system.

For testing this hypothesis, a non-experimental research design with two observations and no treatment was deemed appropriate (Trochim & Donnelly, 2007). First, SFT performance was observed by using research instruments and methods, which have already been applied in prior work. A SFT test was compiled using three rather simple paper-and-pencil tasks. Second, SFM performance was observed by employing a simple inventory management game. The subjects had to place orders for one single product stored in one single inventory with no interconnectedness to other inventories. The orders led to an inflow of products to the inventory after a lead time of four weeks. The outflow was determined by stochastic demand; excess demand was lost. The game was run over 25 periods. The objective was to minimize cumulated costs, which consisted of inventory, ordering, and stock- out costs.

To trial the research design outlined above, a pilot study was conducted. Two classes of a specialization course in operations management with 11 and 19 undergraduate students in their last semester were used as a research site. By the time the tests were carried out, the students should have had a solid knowledge in business administration and quantitative methods. They also had attended two prior courses in operations management; their knowledge in inventory control, however, was basic. The SFT test and the inventory management game were carried out in the very first session of the supply chain management module at the end of September 2008. The game was played first and took about 60 minutes. After a 15 minute break, the SFT test was done second. The participants were to spend up to 30 minutes on that test and, when completed, were allowed to leave. The average time spent was approximately 15 minutes. No incentive was offered, which possibly reduced the participants' motivation and effort. Yet, the effects of financial incentives in experiments have been found to be ambiguous (Camerer & Hogarth, 1999). Instead, the subjects were openly asked to do their best, because the results would be used anonymously in a research project. In the end, 26 students participated, 22 men and 4 women.

DIFFICULTIES IN UNDERSTANDING ACCUMULATION CONFIRMED

For assessing the SFT ability, three relatively simple paper-and-pencil tasks were compiled that had already been used in prior studies in an identical or very similar form (Cronin et al., 2007; Cronin et al., 2009; Kainz & Ossimitz, 2002; Ossimitz, 2002; Sterman, 2002). Each task was designed to measure the subjects' understanding of stocks and flows and their ability to infer their behaviour over time. The type of the tasks ranged from sketching behaviour over time patterns, reading and interpretation of line graphs to multiple choice questions.

The first task is referred to as rainwater tank (RWT) task. It was taken from Kainz and Ossimitz (2002). The instructions were as follows: "A rainwater tank with a capacity of 100 litres is empty at noon. At exactly 2 PM, rain sets in and the water from the gutter flows into the tank at a rate of 25 l/hr until midnight. The tank has no drain; it simply fills up until the water spills over. The spill over is seen as outflow." The task was threefold: The subjects were asked to sketch in a behaviour-over-time chart, which was

provided as template, (1) the inflow of rain water in the tank, (2) the outflow spilling over the tank, and (3) the stock of water in the tank.

The correct solution to this task is provided in Figure 1. Both inflow and outflow step up instantaneously at 2 PM and 6 PM respectively. In-between 2 PM and 6 PM, the water level in the tank rises at the constant rate of 25 litres per hour. After four hours of constant inflow, the volume of water in the tank has accumulated to 100 l, which exhausts the tank's filling capacity. From 6 PM onwards, water spills over at a constant rate of 25 litres per hour.

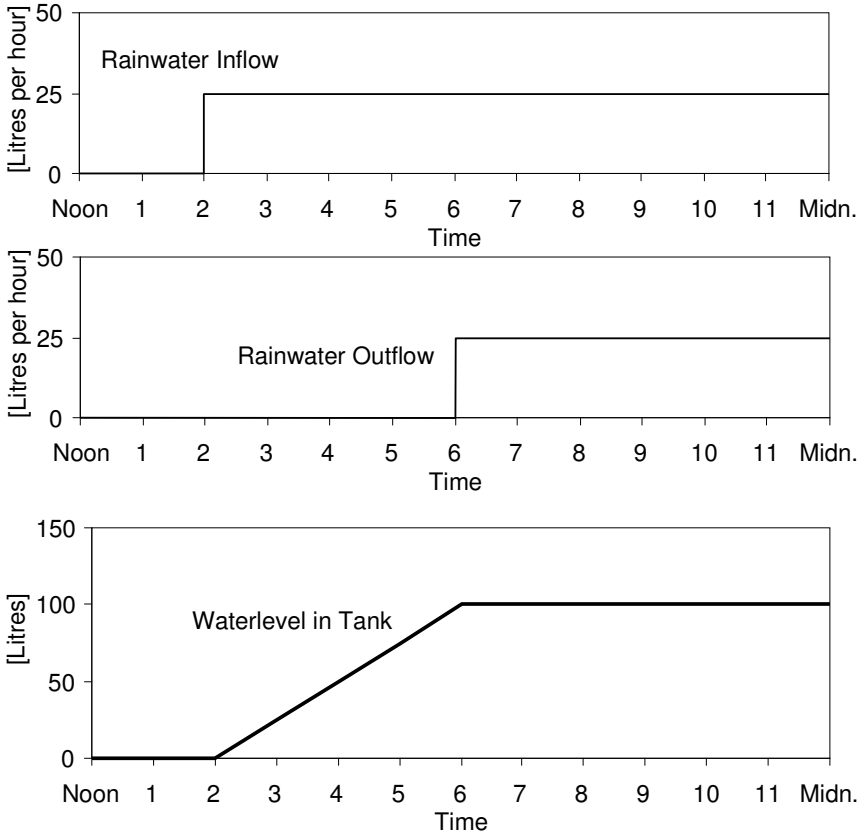

Figure 1: Correct Solution to the RWT Task

As Table 1 shows, the performance in the RWT task was poor. Only five out of 26 students completed all three tasks accurately. While almost everybody succeeded in correctly sketching the behaviour over time of the stock, less than 1/3 charted the inflow and outflow graphs properly. While some of the participants erroneously the same time pattern for the inflow as for the stock, others got the instantaneous step up at 2 PM wrong. Those graphs showed an inflow of zero for 2 PM and then an upward ramping line, which became flat at 25 litres per hour for 3 PM till 12 PM. Not surprisingly, the same ramp-up error could be observed for the outflow. Sometimes, subjects drew a straight line for the outflow ramping up from 0 litres per hour at 6 PM to 150 litres per hour at 12 PM. When compared to the results of Kainz and Ossimitz (2002), this study finds somewhat less disastrous results. Specifically, it is noticeable that all but two students got the time pattern for the stock correct.

	RWT 1	RWT 2	RWT 3	All 3 Tasks	N
This study	30,8%	30,8%	92,3%	19,2%	26
Kainz Ossimitz 2002	na	na	47%	3%	64

Table 1: Percentages of Correct Answers in the RWT Task

The second task was adapted from Cronin et al. (Cronin et al., 2009). While the identical data basis was used, the cover story was slightly changed. Instead of persons entering and leaving a department store, Figure 2 was said to show the number of people entering and leaving a bank branch. This task is therefore referred to as bank branch (BB) task. The questions asked remained unchanged. The first two questions were: "During which minute did the most people enter (leave) the store?" These two questions served the purpose of identifying whether the subjects were able to correctly read line graphs. They do not contribute to the assessment of SFT performance; consequently, these two questions are not included in the aggregate SFT performance measure. As Table 2 indicates, the results do not differ much for the two tasks BB1 and BB2 among the four studies listed. Very high percentages of correct answers suggest that reading line graphs is not the problem.

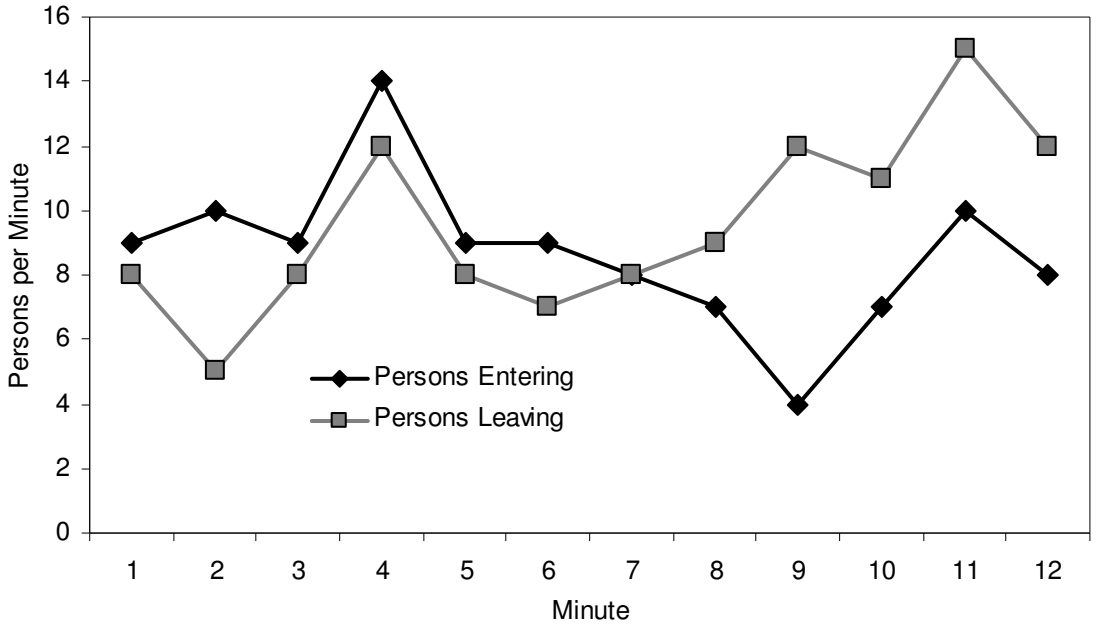

Figure 2: Bank Branch Task

In obvious contrast to this, the high percentages of wrong answers to questions three (BB3) and four (BB4) reflect the erroneous understanding of stocks and flows. Question BB3 was as follows: "During which minute were the most people in the bank branch?," while BB4 asked "During which minute were the fewest people in the store?"

	BB1	BB2	BB3	BB4	BB3 & 4	All 4	N
This study	88,5%	92,3%	30,8%	42,3%	19,2%	15,4%	26
Cronin et al. 2009	94%	87%	52%	41%	na	na	63
Sterman 2002	94%	94%	42%	30%	na	na	172
Ossimitz 2002	na	94%	24%	na	na	na	154

Table 2: Percentages of Correct Answers in the Bank Branch Task (or Comparable Ones)

The determination when the most people were in the branch does not require any calculation, although it would, of course, be possible to find the correct answer by extracting the data points from the line graph, subtracting the number of persons leaving from the number of persons entering per minute, and cumulating the net flow over the time span of twelve minutes. The correct answers can be found more easily and more quickly if one recognizes that the number in the branch accumulates the flow of people entering less the flow of people leaving, and if one understands that a stock rises when its inflow exceeds its outflow and falls when outflow exceeds inflow. Since until minute 7 the number entering always exceeds the number leaving, the stock of people in the branch grows. From minute 8 onwards, the outflow is bigger than the inflow, and therefore the number of persons lingering in the bank branch falls. The most people are in the branch when the inflow curve crosses the outflow line, which happens to be in minute 7. As in the other studies reported in Table 2, answers were considered correct if they were within ± 1 of the correct response, that is, minute 6, 7, or 8. The fact that still less than one third of the subjects in my pilot study managed to provide an accurate answer (and less than 50 % in the other studies) can be regarded as strong evidence of stock-flow thinking failure.

Again, no calculation is required to find the correct answer for BB4. Based on the reasoning portrayed above, one knows that the number of people in the branch rises through minute 7 and falls thereafter. Consequently, the fewest persons are in the bank branch either at the beginning or at the end. Determining which is the case requires an assessment whether the area between the rate of entering and rate of leaving up to minute 7 is greater or smaller than the area between the two curves from minute 8 on. As the second area is, in fact, twice as large as the first area, it is not really difficult to deduce minute 12 as the correct response – if one ever reaches that level of reasoning.

In the studies by Sterman (2002) and Ossimitz (2002), a more complex version of the task was used. Instead of 24 data points, line graphs over 30 minutes were shown, which amounts to 60 data points. However, Cronin et al. (2009) tested the hypothesis that performance would improve in simpler versions of the task with fewer data points and rejected it. When comparing the results for BB3 and BB4 provided in Table 2, the fact must be noted that in this study subjects performed better in BB4 than in BB3, while Cronin et al. (2009) and Sterman (2002) reported a reverse order. Because of the rather small target population in my study, the disparity in performance results observed is probably statistically not significant.

The third and last task intended to test whether the subjects were aware of the difference between the net flow "budget deficit" and the stock "national debt." It is referred to as budget deficit (BD) problem. No graphical presentation of information was given or required in this task. Instead, it consisted of six multiple choice questions, which had to

be answered by checking one of four possible answers. The instructions were as follows: “In Taka Tuka land, the amount by which the annual federal expenses exceed the annual federal income, is defined as budget deficit. In 2006, the budget deficit was 60 billion thaler; a year later it was 40 billion thaler.” The subjects were given six statements, which are, along with the correct answer, shown in Figure 3. As an answer, the participants had to check one of the options “correct”, “wrong,” “not answerable,” and “don’t know.”

		Correct	Wrong	Not answer-able	Don‘t know
BD1	In 2007 20 Billion thaler of public debt have been paid back.	☐	☒	☐	☐
BD2	The minister of finance could reduce the public debt from 2006 to 2007 by a third.	☐	☒	☐	☐
BD3	If the minister of finance is able to reduce the federal budget deficit to zero thaler, (a balanced budget), then Taka Tuka land is free from debt.	☐	☒	☐	☐
BD4	The national debt in Taka Tuka land grew both in 2006 and in 2007.	☒	☐	☐	☐
BD5	If the minister of finance is able to reduce the budget deficit permanently to zero thaler (a balanced budget) and there was no budget surplus in the past, then public debt has reached its highest level.	☒	☐	☐	☐
BD6	A decreasing budget deficit automatically implies a decrease in public debt.	☐	☒	☐	☐

Figure 3: BD Task and Correct Solution

To find the correct solution, one has to see the “budget deficit” as net inflow into the stock “national debt.” Such a simple system with one inflow to one stock can only show three possible behaviour modes over time: First, as long as the inflow is positive, the stock is increasing; second, when the inflow is zero, the stock is staying constant, and third, only if the inflow is negative can the stock fall. While – according to the instructions– the net inflow “budget deficit” to the stock “national debt” is decreasing from 60 to 40 billion thaler, the stock itself is still increasing. Consequently, the statements BD1 and BD2 are wrong. By the end of 2007, national debt amounts to 100 billion plus the unknown opening stock in 2006. If the finance minister were able to achieve a balanced budget, for example, in 2008, the net flow would be zero, and national debt would stay constant at 100 billion. It never ever could be zero itself, which means that BD3 is wrong, too. Assuming that there was never a budget surplus in the past, the reduction of the budget deficit to zero would indeed keep the national debt at its highest level; BD5 is therefore correct. Because of a positive budget deficit for both 2006 and 2007, the stock is increasing in both years; BD4 is correct, too.

	BD1	BD2	BD3	BD4	BD5	BD6	All 6	N
This study	53,8%	50,0%	69,2%	61,5%	46,2%	96,2%	23,1%	26
Ossimitz 2002	68%	36%	62%	44%	17%	42%	na	154

Table 3: Percentages of Correct Answers in the BD Task

Once more, the subjects' performance is poor. Table 3 compares this study's results to the outcomes reported by Ossimitz (2002). Except for BD1, I find consistently higher percentages of correct answers. In particular, BD5 and BD6 stand out with a performance more than twice as high as in Ossimitz's study. Nevertheless, only about 50 % of correct answers in response to the rather simple stock and flow problems BD1, BD2, and BD5 constitute a fairly devastating result.

Overall SFT performance was poor. Only one person out of 26 was able to find the correct solution to all eleven tasks. One subject got only one answer right. On average, about 55 % of the answers given were correct. This study corroborates the findings of previous work (Booth Sweeney et al., 2000; Cronin et al., 2007; Cronin et al., 2009; Kainz et al., 2002; Ossimitz, 2002; Sterman, 2002; Sterman et al., 2002, 2007). Once more, it demonstrates a profound and notable shortcoming in human reasoning: The inability of even smart and well-educated people to understand the dynamic relationships between stocks and flows, that is, the process how flows into and out of a stock accumulate over time. Cronin et al. (2009) demonstrate that poor SFT performance persists regardless of the cover story, the display format of the data, and the quantity of information provided. They reveal that learning is slow when tasks can be done repeatedly and outcome feedback is provided. Moreover, they show that modest incentives do not improve performance. This last finding matches my own experience. In a written exam for the master level, I included the same bank branch task. As this exam was graded, students should have had a sufficient incentive to do as best as they could. However, the outcome did not differ much from the results shown in Table 2. Out of 18 students, only 9 (50 %) got BB3 and only 7 (38 %) got BB4 right.

In dynamic decision making and system dynamics literature it is hypothesized that effective decisions in dynamic settings require decision makers to understand accumulation (Cronin et al., 2009; Dörner, 1996; Pala & Vennix, 2005; Sterman, 2002). Consequently, if SFT performance is as poor as this and various other studies show, decision making performance is expected to be extremely deficient, too. The next section outlines how decision making performance in a dynamic inventory management task is determined. The subsequent chapter then analyses whether SFM performance and SFT performance is correlated to test the hypothesis outlined above.

AN INVENTORY MANAGEMENT GAME FOR DETERMINING SFM PERFORMANCE

Inventory management games were developed many years ago and used as educational instruments in practice and academia for many years (e.g., Renshaw & Heuston, 1957). They have also been used as research laboratories (e.g., Barlas & Özevin, 2004; Sterman, 1989a, b). The inventory management game utilized in this study to measure judgemental decision making performance is regularly played in operations

management courses at both Bachelor and MBA levels.[1] By "practicing" inventory management, students get a realistic impression of the challenges and possible strategies in inventory control problems. To facilitate learning, detail complexity of the task is kept as simple as possible; the same holds true for dynamic complexity. Feedback and non-linearities are not included. This limited complexity also qualifies the game to serve as a measuring device for SFM performance in this study. According to the game's cover story, participants act as inventory expeditors in a retail company. They are responsible for one single product, for example, a flat screen television set. Aware of all relevant costs and the lead time, their only decision is when to reorder how many items so that total costs are minimized. The game is run over a time span of 25 weeks, and therefore participants have to make 25 decisions.

To make the game realistic and minimize the chance that smart participants base their decisions on optimal solutions derived from simple standard inventory management models, a more advanced task with a lead time of four weeks and inventory as well as ordering and stock-out costs is chosen. Inventory costs amount to 2 € per item per week; one order is said to cost 100 €, and stock-out costs per item stand at 70 €. By setting demand for the product randomly beyond the players' control, the possibility of calculating the optimal strategy is eliminated. The players have indeed to rely on their judgement – potentially enhanced by some simple additional calculations.

Demand is determined using a set of playing cards. The number of cards included is shown in Table 4. Additionally, each card reveals the assigned demand; if, for example, in week 8 an ace is drawn, the demand for that week is one item. If a card showing a king is drawn, two additional cards are taken, and the demand for all non-king cards is added up. If once more a king card is unveiled, two additional cards are taken again. If two king cards are put on the table, four more cards are drawn. In all cases, the demand is derived by adding up the figures.

	King	Joker	Ace	2	3	4	5	6	7	8	9	10
Number of cards	3	1	1	2	4	6	8	8	8	8	5	3
Demand	see Figure 4	0	1	2	3	4	5	6	7	8	9	10

Table 4: Playing Cards Used in the Inventory Game

[1] I am grateful to Professor Dr. Rainer Sibbel, who brought the game to my attention and helped me to run it myself.

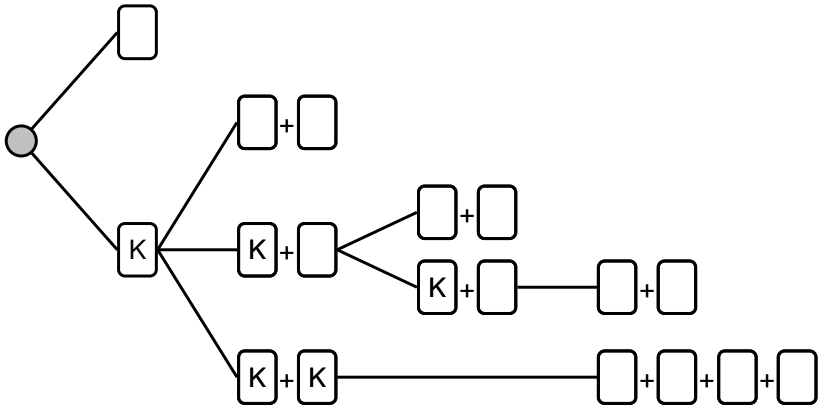

Figure 4: Drawing Algorithm Used in the Inventory Game

By using the set of cards shown in Table 4 and the rather complex algorithm illustrated in Figure 4, determining the probability density function is not straightforward. While it is obvious that minimum demand per week is zero and maximum demand is 39, working out the exact frequency distribution would be impossible in the approximately 15 minute time span, which is provided to develop a procurement policy. Consequently, the participants are forced to rely on their judgements and – depending on the decision rule they choose – on proximate calculations. If necessary, for example, they could approximate the expected demand by ignoring the king playing cards and the complicated drawing algorithm associated with them. This would result in an expected demand of 5.375 items per week.

The game is thoroughly introduced: Participants receive both a verbal briefing and written instructions, which outline the setting, the task, and the objective. A protocol sheet is handed over to the participants. The sheet is blank except for the initial inventory of 40 for week one. Participants receive explanations how to fill in and calculate the cells. Two or three rows are completed as an exercise to ensure as best as possible that the participants do not make mistakes. The game is run in two steps per week. First, demand is determined using the set of playing cards. Second, each player calculates inventory, sales, closing inventory, and shortages. The decision on the order quantity for that week is made and recorded both in column I and – four weeks later – in column D. Finalising step two, inventory, ordering, and shortage costs per week are calculated and entered. When week 25 is finished, the participants add up all costs and calculate total costs. The results are exchanged and discussed in class.

After class, each subject's decision series is transferred to a spreadsheet, which automatically recalculates all measures and cost outcomes. In one case, errors that could not be corrected were discovered; this case had to be eliminated, which resulted in a reduced N of 25.

While the widespread evaluation of decision quality based on outcomes can be criticised (e.g., Davern, Mantena, & Stohr, 2008; Keren & Bruin, 2003), the same laboratory situation for all participants allows to control for factors otherwise beyond control, such as topicality and comparability of information. Following Sterman (1989a; 1989b), Süß (1996), and many others, this study therefore uses the following outcome performance measures: inventory, ordering, stock-out, and total costs. Inherent in the game's design

is that increasing the order quantity also increases inventory costs, yet decreases stock-out costs. More frequent orders increase ordering costs but decrease inventory costs. Consequently, total costs can be seen as a balanced measure for decision quality in the inventory game and is therefore used as measure for inventory management performance. While absolute cost figures could be used to rank the participants of one game session, they neither allow for drawing a comparison between two games based on different demand scenarios nor for making a judgment about the inventory management quality in general. For these reasons, the outcome of an optimized standard (Q, r) inventory control policy[2] is calculated as benchmark. As this benchmark is just used to make participants' results comparable, no attempt is made to find the overall optimal policy. Indeed, any policy could be used to serve as benchmark policy. From an educational perspective, the (Q, r) policy has the advantage that it is rather simple and easily understood by the participants. They readily agree to this benchmark policy and accept its outcome as a point of reference.

The optimal order quantity Q and reorder point r for the given demand probability distribution are determined using a simulation model of the game (see Figure 5) and the Powell optimization procedure built into the software package Vensim. Minimizing expected total costs over 500 different time series of random demand is used as objective function. Q and r are set as parameters, which are to be optimized. The Powell search is restarted 10 times and results in an optimal reorder point r of 35 and an optimal order quantity Q of 28. Expected total costs are 1,973 €, minimum costs amount to 1,616 €, while maximum costs result in 3,978 €.

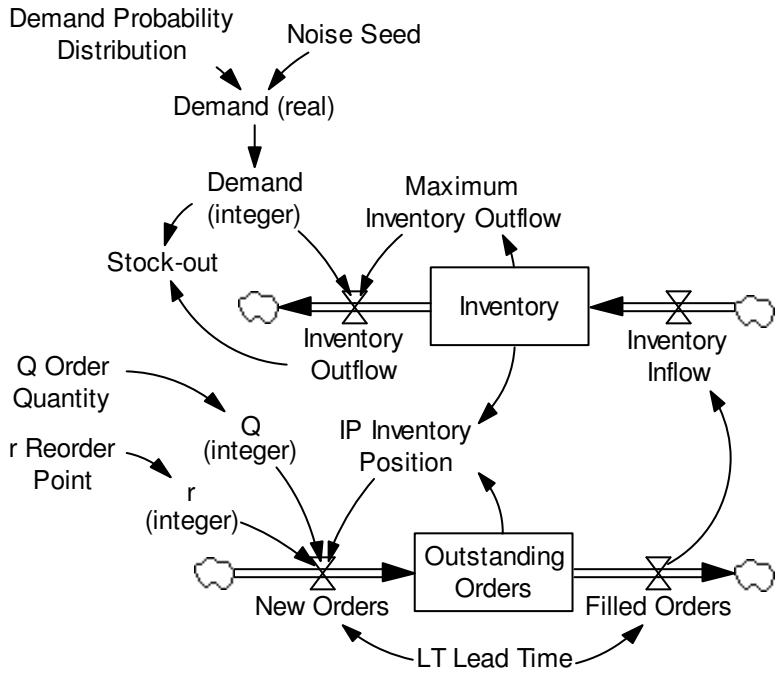

Figure 5: Stock and Flow Structure of the Inventory Game Simulation Model

[2] When the inventory position falls below r, an order for Q is placed.

As the inventory game was played in two different classes, two different demand patterns were generated (see Figure 6, left side). Both demand time series are characterized by at least two outliers resulting in total costs close to the maximum value. Benchmark costs of 3,856 € for class 1 and 3,476 € for class 2 show that the (28, 35) order policy does result in total costs nearly double the expected value (1,973 €). While most participants should be able to outperform the (28, 35) policy, this is no objection to using these cost figures as benchmarks.

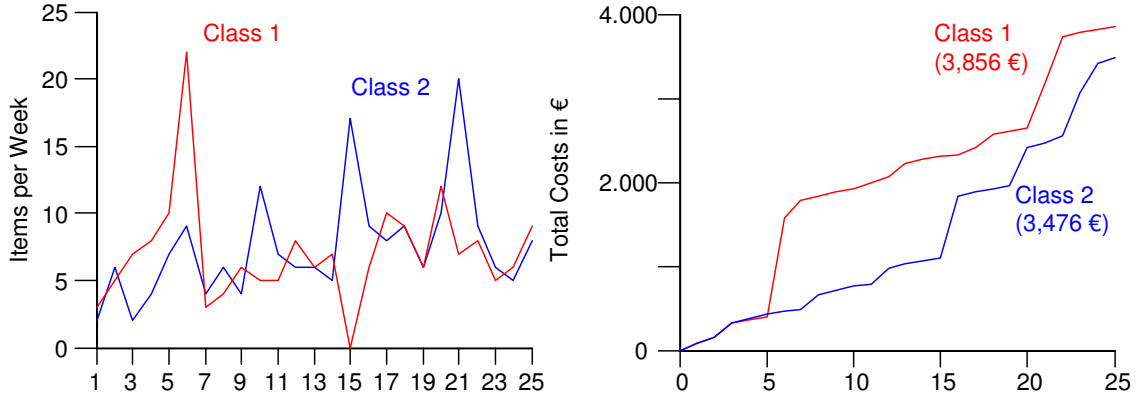

Figure 6: Outcome of the (28, 35) Benchmark Policy

The inventory management performance (IMP) measure is calculated by subtracting the total costs achieved by the participants from the benchmark total costs. For example, a participant in class 1, who obtained total costs of 2,828 €, is assigned 3,856 € – 2,828 € = 1,028 €, which can be interpreted as over-achievement. Therefore, the higher the IMP score the better the performance in IM. Negative values express under-achievement, that is, the participant's performance is below the results of the benchmark policy. As the inventory management game is used as one example of a stock and flow management task, SFM performance is set equal to IMP.

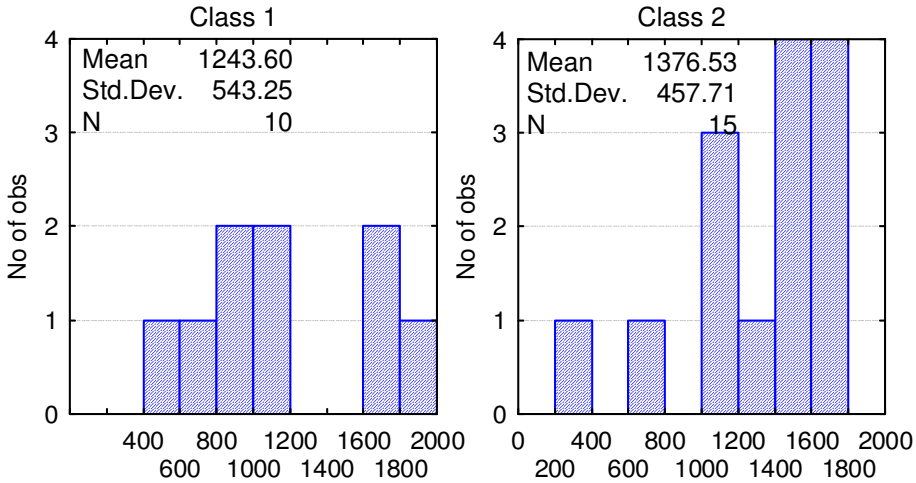

Figure 7: Histogram of the IMP scores

Figure 7 shows the histograms and descriptive statistics for the two classes. Indeed, all participants outperform the (28, 35) benchmark policy. Nevertheless, there would be still room for improvement. When comparing the mean total costs to an ideal ordering policy, class one participants could have achieved 1,548 € instead of a mean of 2,612 €; in class two, minimal total costs could have been 1,566 € instead of a mean value of 2,099 €.[3] One can conclude that the performance of the participants in the SFM task is better than the rather poor results in the SFT test indicate. However, almost all participants could have performed better. Only the top performers come close to the outcome of an ideal inventory management policy.

TESTING THE LINK BETWEEN SFT- AND SFM-PERFORMANCE

The hypothesis that was stated above links the understanding of how flows accumulate in stocks to the performance in managing a stock and flow system. By using the observation settings outlined in the previous chapters, this hypothesis is operationalised as follows:

The better people do in the SFT tasks, that is, the higher their score, the better they perform in the inventory management game, which means the higher their IMP measure (= SFM performance).

Statistically, this hypothesis can be tested by determining the correlation coefficient for the SFT and IMP measures. As both measures are not normally distributed, the nonparametric Spearman Rank Order Correlation Coefficient is calculated.[4] The results for groups one and two separately as well as the whole data set are shown in Table 5. As is obvious from the table, no acceptable p-level can be found. Therefore, the hypothesis has to be considered as rejected.

	Group	Valid N	Spearman R	t(N-2)	p-level
SFT & IMP	1 & 2	25	-0.2080	-1.0201	0.3183
SFT & IMP	1	10	-0.5031	-1.6466	0.1382
SFT & IMP	2	15	0.0615	0.2221	0.8277

Table 5: Spearman Rank Order Correlations for SFT and IMP Total Costs

For group one and the whole data set, even negative (although insignificant) correlations are calculated. This would stand for the exact opposite of the hypothesised relation. As this test outcome is rather unexpected, it might be worthwhile to additionally look at the detailed cost performance measures. Table 6 shows the Spearman Correlation Coefficients for SFT performance and IMP ordering, stock-out, and inventory costs. Again, with one exception, no significant correlations can be found. And the one significant correlation expresses that the higher the participants'

[3] The ideal ordering pattern was determined using the actual demand sequence and optimizing the 25 decisions on the order quantity separately using again Vensim's Powell optimizer. The grid search was restarted more than 100 million times. The participants, however, could have achieved those results only by mere chance; even if they had correctly anticipated demand, it would have been impossible for them to find the optimal solution in the short time span available. Therefore those ideal results were not used as benchmarks.

[4] Missing data are pair-wise deleted.

SFT ability was, the worse was their IMP measured by stock-out costs. Once more, this contradicts the hypothesis stated above.

	Group	Valid N	Spearman R	t(N-2)	p-level
SFT & IMP Ordering Cost	1 & 2	25	0.1966	0.9617	0.3462
SFT & IMP Stock Out Costs	1 & 2	25	-0.2005	-0.9816	0.3365
SFT & IMP Inventory Costs	1 & 2	25	-0.3307	-1.6805	0.1064
SFT & IMP Ordering Cost	1	10	0.3469	1.0461	0.3261
SFT & IMP Stock Out Costs	1	10	-0.6894	-2.6915	**0.0274**
SFT & IMP Inventory Costs	1	10	-0.4260	-1.3316	0.2197
SFT & IMP Ordering Cost	2	15	-0.0577	-0.2083	0.8382
SFT & IMP Stock Out Costs	2	15	0.1701	0.6222	0.5445
SFT & IMP Inventory Costs	2	15	-0.2496	-0.9292	0.3697

Table 6: Spearman Rank Order Correlations for SFT Performance and IMP Detailed Costs

The rather unexpected results of the correlation analysis raise the question why the widely claimed causal relationship between SFT and SFM performance was not found. This could, of course, be attributed to the fact that the sample size in this preliminary study was very small. The number of 25 valid observations is clearly below the N achieved in other studies. However, the scatter plot shown in Figure 8 suggests extending the number of observations in future research and investigating further influencing factors and developing a more complex causal model.

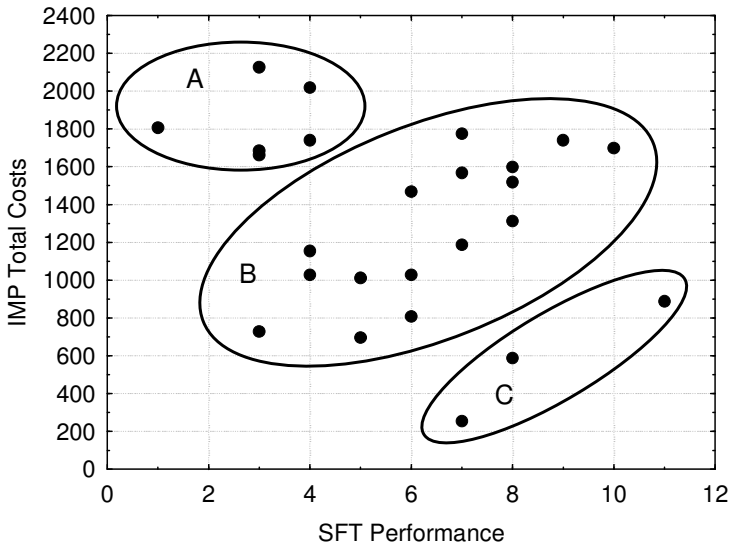

Figure 8: Scatter Plot Visualising the Relation between SFT Performance and IMP Total Costs

At first glance, the spatial distribution of the data points confirms the non-significant, yet slightly negative correlation between SFT and IMP performance. However, after a while, three clusters emerge.[5] Cluster A contains a set of seven data points belonging to participants with rather poor results in the SFT test and extremely good results in the SFM task. The B cluster covers the majority of cases. In this cluster, at least visually, the hypothesised positive correlation between SFT and IMP performance holds. The

[5] I am grateful to Professor Dr. Olaf Stotz, who drew my attention to this clustering phenomenon.

third cluster, entitled "C" cluster, encompasses only three participants. They are characterized by rather good results in the SFT test combined with poor results in the inventory management game. Again, the hypothesised correlation seems to hold, yet the extremely low number of three cases does not indicate high resilience. Obviously, the reliability and robustness of these results are limited by the rather low number of in total 25 cases.

The three clusters that can be identified in the scatter plot in Figure 8 could indicate that the stated monocausal relationship between SFT and SFM performance is too simple and not appropriate. A more complex causal model with further moderating and directly influencing factors might be needed. In the following section, a first attempt is made to introduce some findings and ideas from psychological research on decision making performance in dynamically complex environments.

TOWARDS A MORE COMPLEX EXPLANATORY MODEL FOR INVENTORY MANAGEMENT PERFORMANCE

Stock and flow management, as operationalised in the game used in this study, is not a dynamically complex task. The inventory system to be managed has no feedback, no non-linearities, and no internal dynamics. It is a simple one stock, one inflow, and one outflow system with a four-week delay between order placement and delivery. Contrary to this, dynamic decision making research focuses usually on far more complex dynamic systems, such as an air traffic control system (Ackerman et al., 1993; Ackerman, Kanfer, & Goff, 1995), a city (Dörner, Kreuzig, & Reither, 1994), or a high tech company (Wittmann & Hattrup, 2004). For these dynamically complex tasks, elaborate and corroborated theories exist that relate intelligence to decision making performance. Especially Ackerman's (1996) PPIK theory has been bolstered by many empirical studies (see, e.g., Wittmann et al., 2004). Despite the much lower complexity of the inventory management task used in this study, building on the PPIK theory as theoretical framework for an advanced follow-up investigation could be a promising approach. Therefore, the PPIK theory is briefly introduced within the following paragraphs, and a more elaborate causal explanation model of SFM performance is suggested.

The first 'P' of PPIK stands for intelligence-as-process, which encompasses reasoning, memory-span (also short-term or working memory), perceptual speed, and spatial rotation (Ackerman, 1996). The second 'P' denotes personality, which is described to include first openness (or similarly defined traits, such as Intellectence or Culture) and second typical intellectual engagement (TIE) (Ackerman, 1996). The 'I' in PPIK represents interests, specifically realistic, investigative, and artistic interests. Finally, the 'K' stands for intelligence-as-knowledge, about which Ackerman (1996) remarks that it has to be seen as contextual.

Based upon Ackerman's PPIK theory outlined above, Figure 9 illustrates an advanced causal model for predicting SFM performance (SFMP). SFT ability, as measured by using the rain water tank task, the bank branch task, and the budget deficit task, could be seen as one component of intelligence-as-knowledge required to support the inventory management game. Of course, other knowledge aspects might be necessary for a satisfactory explanation of SFMP. Therefore, one could try to measure the general

economic knowledge, as, for example, Wittmann and Hattrup (2004) did in their study, as well as specific inventory management knowledge. For the constructs intelligence-as-process, interests, and personality well-tried instruments are available. Intelligenc-as-process could, for example, be assessed using the BIS test (Jäger, Süß, & Beauducel, 1997); for the measurement of personality, the revised NEO personality inventory (Costa & McCrae, 1992) could be used in combination with the inventory developed by Goff & Ackerman (1992). Finally, to measure interests, one could, for example, employ the ACT's Interest Inventory (UNIACT). Besides the inventory management game other SFM tasks could be used to asses SFMP; for example, a SFM task with a more continuous character and less (or no) stochastic elements could provide an alternative measure for SFMP.

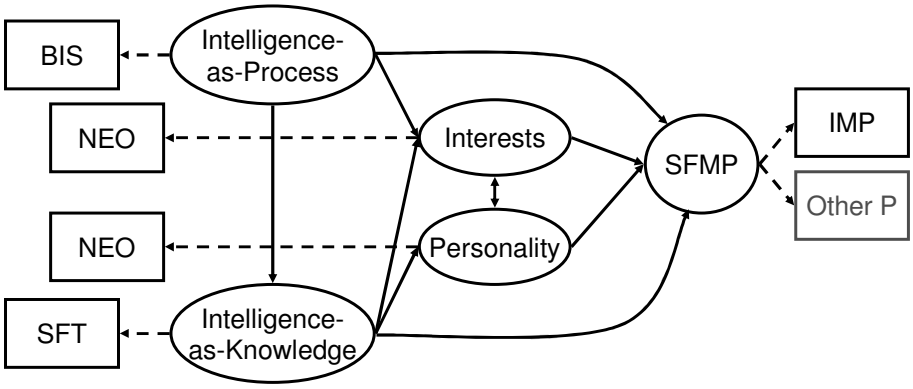

Figure 9: An Advanced Model for Prediction of IMP based on the PPIK Theory

The gains from using established psychological tests could be threefold: First, one could expect to find the clustering factor or factors, which causes the pattern in Figure 8; second, one could investigate how strong SFT ability relates to SFM performance in direct comparison to a selection of other factors already proven to be significant; third; one could build on a vast body of psychological research and learn from it. However, the major disadvantage was that the time exposure of the subjects, which had to be observed in the various tests, would be enormous – approximately four hours.

To reduce the test load for the student subjects, a broad range of factors related to the model depicted in Figure 9 could be assessed in a second pilot study using already existing databases of demographic and study performance data. Where those objective sources are not available, self-assessment questionnaires could help to identify factors with considerable impact. For example, if a university administers an admission test, components measuring intelligence-as-process are often included; these results, although often determined at an earlier point in time, could be used as preliminary indicators. Additionally, grades on various knowledge domains could be analyzed and checked for relevance. Finally, university databases contain a broad range of socio-demographic data, which could be queried to find clustering factors. By doing this, the enormous range of influencing factors on SFMP, which could be derived from the PPIK theory, can hopefully be condensed.

DISCUSSION, LIMITATIONS AND CONCLUSIONS

Initiated and influenced by prior research on stock and flow thinking capabilities, this study focused on the issue whether poor SFT performance, that is, bad understanding of accumulation, indeed predicts poor performance in a dynamic SFM problem. The hypothesis that this causal link does exist can be derived from discussions by, for example, Cronin et al. (2009), Sterman (2002), or Booth Sweeney & Sterman (2000). Those studies, however, primarily attempt to measure SFT performance and try to provide explanations for the poor results – for example, the widespread use of a simple correlation heuristic (Cronin et al., 2009). No attempt is made to test the hypothesised relation empirically. Specifically, this research aims at contributing to narrowing that research gap. The hypothesis that the better people understand stocks and flows, the better they perform in managing an inventory system was empirically tested by setting up a non-experimental research design with two laboratory observations. Using the indeed rather small number of students attending an Operations Management course as a sample, the subjects' SFT performance was determined first, and, in a second step, their success level in a dynamic inventory management task was ascertained. A negative, yet non-significant, Spearman Correlation Coefficient for the total group (N = 25) suggests that the hypothesis has to be rejected.

However, further analysis of the SFT IMP scatter plot revealed three different clusters. For a small group of seven subjects (cluster A), rather low SFT performance came along with high IMP scores; an even smaller group of three subjects, termed cluster C, showed SFT scores above average, yet substandard IMP performance; and for the majority of 15 subjects, assigned to cluster B, average SFT performance was associated with average IMP performance. This clustering can be seen as an indication of the existence of other factors, which were not controlled in the pilot study. If a causal or moderating impact of another factor had been hypothesized and this factor had been controlled for, it is likely that a positive correlation between SFT and IMP performance could have been found. At least the visual impression suggests a strong positive correlation between SFT and IMP performance within clusters B and C. Although no direct evidence for a positive link between SFT understanding and IMP performance can be derived from the pilot study, its outcome is conducive to further research. The results strongly suggest searching for other relevant factors, which cause the clustering and therefore mask a positive correlation within each cluster.

As primarily psychological research (e.g., Ackerman et al., 1993; Wittmann et al., 2004) has already tried to predict dynamic decision making performance using various intelligence constructs, the search for the other factors could benefit from their findings. Therefore, it is suggested to employ Ackerman's established PPIK theory as a framework guiding the development of more complex causal models explaining SFM performance. The main contribution of this pilot study can indeed be seen in providing valuable information supporting an advanced research design for a follow-up study. While one major limitation of this study is indeed the low number of cases (N = 25), simply increasing the sample size without advancing the causal model probably will not yield better results. Future research should try to control a broad range of other factors presumably influencing or moderating the subject's success in the SFM task. As a consequence, the need may arise to add more observations to the non-experimental laboratory research design.

The pilot study, however, does not give rise to a fundamental alteration of the design. The inventory used for measuring the SFT performance seems to be appropriate. The poor understanding of accumulation found by others (Cronin et al., 2009; Kainz et al., 2002; Ossimitz, 2002; Sterman, 2002) could be confirmed. Although the inventory game was not used in prior research to determine SFM ability, no obvious shortcomings were detected while conducting the study and analysing its outcomes. It was rather easy to conduct and produced sufficiently strewing performance measures. While one might argue that the stochastic nature of the task adds another unwanted source of disturbance, a stochastic demand pattern seems to be the only way to provide for the formation of multiple groups without running the risk that information about demand is spread among the groups. Stochastic demand also allows to replicate the game, for example, to determine learning. Yet, for a follow-up study, the participants' stochastical knowledge could be measured, too.

This research engaged in a first attempt to determine the contribution of SFT ability to predict decision making performance in IM. Further research should help to build and test a more elaborate explanatory model. A better understanding is needed how SFT ability fits in the psychological intelligence construct and contributes to dynamic decision making performance. Based on such a theory, improved methods could be developed to, firstly, educate all of us to reach higher levels of performance in stock and flow problems and, secondly, to select the best people for the most demanding stock and flow management tasks.

REFERENCES

2008a. "Out of stock" tops list of shoppers' pet hates. ***In-Store***: 6-6.

2008b. Retailers running risk of abandoned shopping trolleys. ***In-Store***: 7-7.

Ackerman, P. L. 1996. A theory of adult intellectual development: Process, personality, interests, and knowledge. ***Intelligence***, 22(2): 227-257.

Ackerman, P. L., & Goff, M. 1994. Typical Intellectual Engagement and Personality: Reply to Rocklin (1994). ***Journal of Educational Psychology***, 86(1): 150-153.

Ackerman, P. L., & Heggestad, E. D. 1997. Intelligence, Personality, and Interests: Evidence for Overlapping Traits. ***Psychological Bulletin***, 121(2): 219-245.

Ackerman, P. L., & Kanfer, R. 1993. Integrating Laboratory and Field Study for Improving Selection: Development of a Battery for Predicting Air Traffic Controller Success. ***Journal of Applied Psychology***, 78(3): 413-432.

Ackerman, P. L., Kanfer, R., & Goff, M. 1995. Cognitive and Noncognitive Determinants and Consequences of Complex Skill Acquisition. ***Journal of Experimental Psychology: Applied***, 1(4): 270-304.

Barlas, Y., & Özevin, M. G. 2004. Analysis of stock management gaming experiments and alternative ordering formulations. ***Systems Research and Behavioral Science***, 21(4): 439-470.

Bloomfield, R. J., Gino, F., & Kulp, S. 2007. Behavioral Causes of the Bullwhip Effect in a Single Echelon: SSRN.

Booth Sweeney, L., & Sterman, J. D. 2000. Bathtub Dynamics: Initial Results of a Systems Thinking Inventory. ***System Dynamics Review***, 16(4): 249-294.

Brehmer, B. 1995. Dynamic decision making: human control of complex systems. ***Acta Psychologica***, 81: 211-241.

Camerer, C. F., & Hogarth, R. M. 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.

Costa, P. T., & McCrae, R. R. 1992. ***Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI)***. Odessa: Psychological Assessment Resources.

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

Cronin, M. A., Gonzalez, C., & Sterman, J. D. 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.

Croson, R., & Donohue, K. 2003. Impact of POS Data Sharing on Supply Chain Management: An Experimental Study. ***Production and Operations Management***, 12(1): 1-11.

Croson, R., & Donohue, K. 2006. Behavioral Causes of the Bullwhip Effect and the Observed Value of Inventory Information. ***Management Science***, 52(3): 323-336.

Davern, M. J., Mantena, R., & Stohr, E. A. 2008. Diagnosing decision quality. ***Decision Support Systems***, 45(1): 123-139.

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

Dörner, D. 1996. ***The logic of failure. Strategic thinking for complex situations***. New York: Metropolitan Books.

Dörner, D., Kreuzig, H. W., & Reither, F. 1994. ***Lohhausen. Vom Umgang mit Unbestimmtheit und Komplexität*** (Unveränd. Nachdr. der Ausg. von 1983 / mit einem Beitrag von Dietrich Dörner Zwölf Jahre danach: Lohhausen im Rückblick. ed.). Bern et al.: Huber.

Foster, L. 2004. Excess inventory stalls Sears performance, ***Financial Times***.

Goff, M., & Ackerman, P. L. 1992. Personality-Intelligence Relations: Assessment of Typical Intellectual Engagement. ***Journal of Educational Psychology***, 84(4): 537-552.

Hansell, S. 2008. The iPhone Shortage, ***The New York Times***.

Hastie, R., & Dawes, R. M. 2001. ***Rational choice in an uncertain world. The psychology of judgement and decision making***. Thousand Oaks, Calif. ; London: Sage Publications.

Jäger, A. O., Süß, H.-M., & Beauducel, A. 1997. ***Berliner Intelligenzstruktur-Test. BIS-Test. Form 4.*** . Göttingen et al.: Hogrefe.

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

Kainz, D., & Ossimitz, G. 2002. ***Can Students Learn Stock-Flow-Thinking? An Empirical Investigation***. Paper presented at the 20th International Conference of the System Dynamics Society, Palermo.

Keren, G., & Bruin, W. B. d. 2003. On the Assessment of Decision Quality: Considerations Regarding Utility, Conflict and Accountability. In D. Hardman, & L. Macchi (Eds.), ***Thinking: Psychological Perspectives on Reasoning, Judgment and Decision Making***: 347-363.

Kleinmuntz, D. N. 1985. Cognitive Heuristics and Feedback in a Dynamic Decision Environment. ***Management Science***, 31(6): 680-702.

Milne, R. 2005. Bosch in move to put pump problems behind it, ***Financial Times***.

Ossimitz, G. 2002. ***Stock-flow-thinking and reading stock-flow-related graphs: An emperical investigation in dynamic thinking abilities***. Paper presented at the 20th International Conference of the System Dynamics Society, Palermo, Italy.

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.

Reichert, U., & Dörner, D. 1988. Heurismen beim Umgang mit einem "einfachen" dynamischen System. ***Sprache und Kognition***, 7(1): 12-24.

Renshaw, J. R., & Heuston, A. 1957. The Game Monopologs, ***RAND Research Memorandum***.

Senge, P. M. 1990. ***The fifth discipline. The art and practice of the learning organization***. New York: Doubleday.

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(4): 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.

Süß, H.-M. 1996. ***Intelligenz, Wissen und Problemlösen. Kognitive Voraussetzungen für erfolgreiches Handeln bei computersimulierten Problemen***. Göttingen et al.: Hogrefe.

Trochim, W. M. K., & Donnelly, J. P. 2007. ***The Research Methods Knowledge Base*** (3 ed.): Atomic Dog Publishing.

Wittmann, W. W., & Hattrup, K. 2004. The relationship between performance in dynamic systems and intelligence. ***Systems Research and Behavioral Science***, 21(4): 393-409.