Tangible Stock/Flow Experiments − Addressing Issues of Naturalistic Decision Making

Andreas Größler & Jürgen Strohhecker

Institute for Management Research, Radboud University Nijmegen, NL
Frankfurt School of Finance and Management, Frankfurt, D

ABSTRACT

The purpose of this paper is to investigate whether stock/flow failures persist in naturalistic decision making environments. A tangible stock/flow experiment is used in this study, which asks participants to pour a certain amount of water into a glass through a funnel in an as short time as possible. Findings are that people on average do not perform better in a tangible stock/flow task than in a computerized or paper-based test of a comparable task. In addition, individual performance in the tangible task cannot be related to performance in a similar paper-pencil stock/flow task. An implication of this study is that naturalistic stock/flow tasks are as difficult for humans to control as more abstract tasks. Further research should address individual differences between the two modes of task (tangible vs. paper-based). A limitation of this study is the usage of one tangible stock/flow task only. The value of this paper lies in the combination of a standard test with a tangible experiment addressing the same cognitive capabilities.

Key words: understanding of accumulation, stock/flow failure, experiment, naturalistic decision making

THE CHALLENGE OF NATURALISTIC DECISION MAKING

Understanding of accumulation (UoA), i.e. knowing about the nature of stocks and flows, is of utmost relevance for a broad range of decision making situations in society, business, and personal affairs. Yet, dynamic decision making research has shown over and over again that humans are not sufficiently capable to understand the difference between stocks and flows, deduce a system's behaviour resulting from the existence of stocks and flows, or control a stock/flow system. However, the way in which results of dynamic decision making have been obtained has inspired criticism from proponents of naturalistic decision making. Their argument is that the artificial situation (in form of computer-based or paper-pencil tests of understanding of accumulation) in which stock/flow failures have been reported biases these outcomes; in naturalistic settings people were very well able to perform accumulation tasks successfully.

This study attempts to contribute to the understanding of stock/flow failure and to dynamic decision making research in general by investigating whether the criticism by naturalistic decision making is substantial. Thus, the purpose of this paper is to find out whether stock/flow failures persist in naturalistic decision making environments. For this purpose, a tangible stock/flow experiment is used in this study, which asks participants to pour a certain volume of water into a glass through a funnel in an as short time as possible. With this experimental set-up we strive to put participants into a naturalistic decision making situation that is gradually more complicated than just filling water into a glass. Furthermore, since we conduct a standard, paper-based understanding of accumulation test in the same study, we are able to compare the performance of humans in naturalistic and in artificial decision contexts employing a stock/flow task.

With this study, we follow a call by Sengupta & Abdel-Hamid (1993, p. 426) to conduct research on the performance of dynamic tasks in naturalistic settings. The scientific relevance of this paper lies in its contribution to the debate about naturalistic vs. dynamic decision making. The practical relevance of the paper is given by the fact that it can help to answer the question if it would "only" require to design appropriate decision making environments to improve decision making quality (as naturalistic decision making ultimately implies) or whether substantial limitations on the individual level would remain even in highly beneficial decision making contexts.

The structure of this paper is as follows. In the next section, we review the relevant literature on dynamic decision making in general and on stock/flow failure in particular. Furthermore, we discuss the criticism by proponents of naturalistic decision making and present the responses from system dynamics researchers. In the section thereafter, the experimental setting is described. Section 4 comprises a discussion of the results of the experiment. The paper closes with a general discussion of implications, limitation of the study and some suggestions for further research.

STOCK/FLOW FAILURES AND RESPONSES TO CLAIMS FROM NATURALISTIC DECISION MAKING

Human decision making performance in dynamic tasks is generally low—compared to absolute and optimal standards as well as in relative terms compared to heuristic best practice strategies. Dynamic decision making research (Brehmer, 1992; Edwards, 1962) has resulted in ample evidence of decision making failures in dynamic complex systems. Such systems consist of stocks and flows and interrelated information links (Forrester, 1961). They are characterized by feedback and delays between cause and effect (Sterman, 1994). Human decision makers perceive these systems often as opaque, incomprehensible and hard to control (Dörner, 1996). On average, people perform rather poorly in managing such systems. This finding persists over a wide range of systems. For example, Dörner and colleagues found lamentable results of participants who were asked to act as mayor of the virtual small town "Lohhausen" (Dörner, 1980; Dörner *et al.*, 1994). Reichert and Dörner (1988) report failures when participants are charged with the task of manually controlling the temperature of a refrigerated warehouse. Sterman (1989) finds average team costs ten times greater than the benchmark using the well-known beer game as an experimental device. In a new product management task, a naïve benchmark policy outperformed the subjects in 87 % of the cases (Paich & Sterman, 1993). Confronted with the challenge to manage a virgin fish stock, 74 % of the participants overinvested in vessels resulting in a worse-than-optimal achievement of the overall target (Moxnes, 1998). Wittman and Hattrup (2004) report widely varying performance of subjects acting as managers of a tailor's shop, a coal-fired power plant and a high-technology company with a range of substituting products to develop and bring to the markets. Recently, Moxnes & Jensen (2009) report on a significant average overshoot of 86 % of an explicit goal of 0.8 g/l in an alcohol simulator experiment. While a well-developed universal theory of dynamic decision making has not yet emerged, the various research efforts over more than two decades have contributed to a better understanding of the "logic of failure," as Dörner (1996) pithily named these phenomena.

Recent research has identified a potential explanatory factor for poor dynamic decision making performance—misunderstanding of accumulation (MoA). The seminal study of Booth Sweeney and Sterman (2000) has revealed that a large fraction of highly educated people is unable to infer the behaviour of even the simplest stock-flow-systems consisting of only one stock, one inflow, and one outflow. 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' dynamics is far from good. 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 conjecture that the misunderstanding of accumulation is a persistent phenomenon, for instance comparable to the deep-rooted problems people have in probabilistic judgements and decision making (Hastie & Dawes, 2001; Kahneman & Tversky, 1972).

Misunderstanding of accumulation and other results of dynamic decision making research have been criticised by proponents of naturalistic decision making (Klein,

2008; Lipshitz *et al.*, 2006; Lipshitz *et al.*, 2001; Zsambok & Klein, 1997). Their main argument is that apparent failures of humans to deal adequately with dynamic complexity do not result from erroneous thinking but are artefacts of the experimental method employed in this research. They claim that people in everyday life are very well equipped to survive and deal with a wide range of situations successfully. Naturalistic decision making argues that those heuristics that have been considered leading to inferior results in dynamic decision making are actually very functional in realistic, natural situations outside the laboratory. This criticism has been countered by, for instance, Booth Sweeney & Sterman (2000) and Sterman&Booth Sweeney (2002) who agreed that people might be able to control simple naturalistic decision making tasks. However, these two authors claim that complex dynamic decisions of our modern times usually are to be made in situations that are very similar to the experimental settings used in their studies: people have to decide on highly abstract decision making situations that are remote from the tangible quality of the type of naturalistic decision making tasks. This, their argument goes, gives the usual dynamic decision making experiments high external validity and renders the criticism of naturalistic decision making irrelevant.

In this study, we want to contribute to this discussion by shedding light on two assumptions in this debate: (i) people are good decision makers in naturalistic situations when it comes to stock/flow tasks and (ii) performance in naturalistic and in more abstract tasks is related. We do not address Sterman and Booth Sweeney's argument that today's decision making contexts rather require control over abstract stock/flow tasks than naturalistic ones. Scrutinizing this argument would require the classification and evaluation of a wide range of decision making tasks that people are confronted with in their professions and daily lives.

The first assumption of the naturalistic vs. dynamic decision making debate that we want to challenge is that people are convincingly able to control naturalistic stock/flow tasks. At least when the level of complexity of these tasks exceeds the simplest possibility (directly filling or draining one stock), the answer to the question whether they can do this well or not is not trivial. While we stick to the experimental method in our study, we use a tangible task that participants have to fulfil: filling a glass with water through a funnel (in structural terms: filling a stock with a delay). Thus, our first proposition is:

P1: Participants are able to achieve good performance (that is, a minor deviation from the target in a reasonable short time span) when conducting a tangible stock/flow experiment; more specifically, when filling a glass with water through a funnel.

Finding support for this proposition would strengthen the argument of the proponents of naturalistic decision making research. Even more support (or refutation), however, would come from a direct comparison of a tangible task with abstract tasks. Therefore, we propose, secondly:

P2: Participants are able to perform better in a tangible stock/flow experiment than
a) in a similar simulator based experiment;
b) in paper-pencil stock/flow tasks.

Finding support for these propositions would give an indication that naturalistic tasks allow for a much better stock/flow performance than abstract tasks since in abstract tasks participants' performance usually is substantially below reasonable benchmarks. Rejecting the proposition would indicate that people do not benefit from naturalistic situations in stock/flow tasks or only when the tasks are extremely simple (i.e. simpler than our task).

Besides a comparison on the aggregate level, comparing individual performance could provide insights in the possibility to generalize behaviour and performance from one task mode to the other. Thus, thirdly, we want to challenge the assumption that performance in naturalistic and in abstract dynamic decision making tasks can be related to each other. The question behind this challenge is whether individual performance in naturalistic stock/flow tasks tells us anything about individual performance in a more abstract task. Therefore, we formulate the third proposition as

P3: Participants performing well in the tangible stock/flow task will also show a good understanding of accumulation in a paper-pencil test and vice versa.

If we found support for this proposition it would give an indication that both tasks (physically filling a glass of water and solving paper-based tests about the accumulation of stocks) are comparable: people performing well in one setting would also perform well in the other. If this proposition were to be rejected it would show that these tasks actually require two different sets of competencies and are difficult to compare directly: the debate would shift from a methodological discussion about the influence of the experimental method on the results into a content debate about what are the psychological foundations of decision making in different contexts.

EXPERIMENTAL DESIGN AND IMPLEMENTATION

Controlled observations in a laboratory are used in this study for gathering data. As the research objective is primarily explorative and descriptive, a non-experimental (correlational or observational) research design is chosen. First, UoA is assessed using a specific inventory that compiles a number of rather simple paper-and-pencil tasks developed by Booth Sweeney and Sterman (2000), Sterman (2002), and Ossimitz (2002). Second, the participants' performance in a tangible decision making setting—filling a glass of water through a funnel—is repeatedly (twice) observed. Although this setting is not a true experiment involving a treatment and a control group, the term experiment is kept nevertheless to describe the process of conducting these observations under controlled conditions.

The laboratory was installed in two adjacent seminar rooms. The larger one for up to 40 persons was equipped with cubicles to minimize interactions between participants. The smaller one was specifically setup to conduct the funnel and glass experiment (as shown in Figure 1). Between the two rooms a participants' waiting area was established.

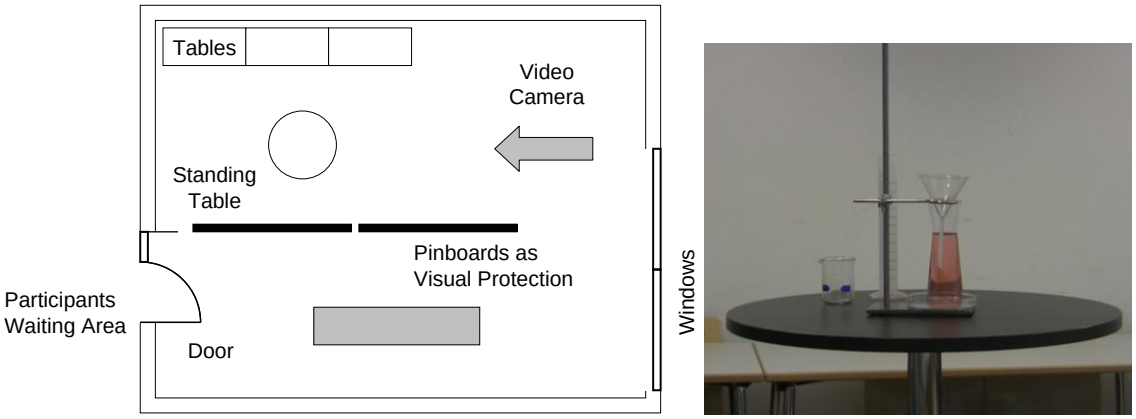
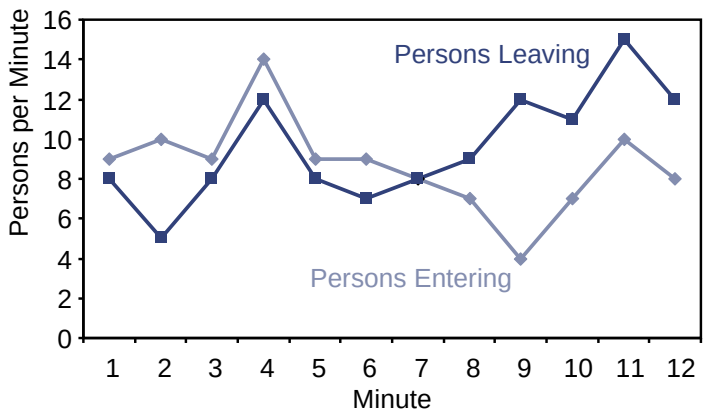


Figure 1: Laboratory setup of the funnel and glass experiment

For assessing UoA ability, five relatively simple paper-and-pencil tasks were compiled that have already been used in prior studies in an identical or very similar form. Each task was designed to measure participants' understanding of stocks and flows and their ability to infer systems 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 was taken from Kainz and Ossimitz (2002) and is referred to as a rainwater tank (RWT) task. The second task was adapted from the department store task developed by Sterman (2002). It was renamed bank branch task (BB); Figure 2 shows this task an illustrative example. The third task intends to test whether the participants are aware of the difference between the net flow "budget deficit" and the stock "national debt." It was adapted from Ossimitz (2002) and is referred to as a budget deficit (BD) problem. Task number four and five were taken from Booth Sweeney & Sterman (2000). The fourth task is the so-called manufacturing case (MC). In this task, participants have to determine a possible reaction of a production facility to a sudden drop of final goods inventory, incorporating a four week delay from starting production to putting final goods on stock. The fifth and last task in the UoA inventory is the bath tub (BT) task.



During which minute were the most people in the bank branch?
During which minute were the fewest people in the bank branch?

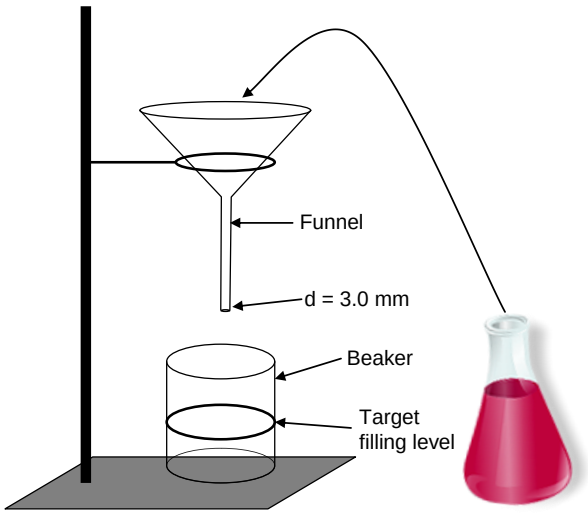
Figure 2: Illustration of the bank branch UoA task

The task used in the tangible decision making experiment was inspired by Moxnes & Jensen (2009) and Moxnes (2011). Moxnes & Jensen (2009) report on the effects of delayed absorption of alcohol and the problems participants have in understanding this lagged process. In a computerized simulator experiment with high school students, who made drinking decisions, the authors observed an average overshoot of 86 % of an explicit goal of 0.8 g/l with a stomach delay time set to 22 minutes, and an average overshoot of 21 % with a stomach delay time of 4 minutes. In both papers the analogy of filling a glass through a funnel is used to explain the observed (and more general) overshoot behaviour. Compared to filling just a glass of water (one stock) using a funnel introduces a second stock that is delaying the inflow of water into the glass. Simple filling heuristics that rely exclusively on feedback about the water level in the glass and—erroneously—ignore the water in the funnel, lead to overshoot (and sometimes to overflow). When one stops pouring once the target level of water in the glass is reached—which is good strategy for just filling a glass—the water will continue flowing from the funnel in the glass. As result, the water level overshoots the target. Moxnes & Jensen (2009) argue that the stomach in their alcohol simulator experiment corresponds to the funnel—with one major difference: while the water in the funnel is visible the alcohol in the stomach is not.

As this study's objective is to investigate whether stock/flow failures persist in naturalistic decision making environments, real objects should be used and all stocks and flows should be clearly visible. Therefore, we take Moxnes & Jensen's (2009) funnel and glass analogy and transform it in a tangible experimental setup. We use a funnel made of glass and a glass beaker. The water was coloured with red ink to improve observability. The target filling level (98 ml) is clearly indicated on the beaker. Participants are assigned the task to fill the beaker up to the indicated target level by pouring water into the funnel in as less time as possible. Precisely, the task in our study was introduced to the participants as described in Figure 33.

Instructions

Please look carefully at the following experiment setup:



It is your task to pour the red liquid from the flask into the funnel so that it flows in the beaker positioned below. You must by all means avoid that the liquid brims over the top of the funnel. You are allowed two runs. Please achieve the following targets as best as you can:

1. Minimize the variance between the target filling level marked on the beaker and the actual filling level (measured in milliliter)!
2. Minimize the filling time, which is measured as the time span between lifting the flask from the table and posing it back onto the table!

For your participation in this experiment you are rewarded depending on the achievement of these targets: The less the volume variance and the less time you need, the higher is the monetary reward. It becomes zero if the liquid brims over the top of the funnel. The reward per run is calculated exactly as follows (negative amounts are set equal to zero):

$$\text{Reward} = 5[€] - 0{,}2\left[\frac{€}{\text{ml}}\right] \cdot \left|\text{Volume Variance}[\text{ml}]\right| - (\text{Filling Time}[\text{s}] - 5[\text{s}]) \cdot 0{,}2\left[\frac{€}{\text{s}}\right]$$

Figure 3: Instructions (authors translation from German task description)

Following suggestions from experimental economics (e.g. Friedman *et al.*, 2004; Guala, 2005), Smith's (1976, 1982) induced value theory is applied and participants are incentivized by a monetary reward. The financial incentive was linked to a participant's performance in both tasks. For the UoA test, the cash-out was calculated according to the percentage of correct answers with a maximum of 10 €. For the funnel and glass task the precise payoff function was provided in the instructions and can be found in Figure 3. On average, 9.04 € were achieved by the participants for an exercise of about 60 minutes.

For reasons of availability and financial feasibility students were used as participants in the experiment. This has also the advantage that the results can be compared to previous studies that have mostly also relied on students. The experimental sessions were integrated into a core course on "Production Management" that is part of a Bachelor of Science Business Administration program at a German business school. In June 2011,

four experimental sessions were performed, involving as participants in total 71 students in their fourth semester. The students had been informed about these special sessions four weeks before the specific dates. Invitations were made both orally in class and per email.

Upon arrival in the larger room, participants were asked to sit down using the cubicles provided and get prepared with a pen. Once the session was started, the participants were briefly introduced to the process. First, the UoA test was handed out and the participants were asked to note start and end time on the cover sheet. Once a participant had completed this test, the experimenter provided the instruction sheet for the funnel and glass test and asked to read through the instructions thoroughly. Having finished his or her reading the participant left the room and waited in the waiting area before the "funnel and glass laboratory". To ensure that participants in the waiting area communicated as little as possible a third supervisor (in addition to two experimenters) was installed.

Before the funnel and glass experiment was started, each participant was asked by the experimenter to pose any remaining questions. All questions were answered and if any confusion existed it was cleared up. The participants were allowed to get prepared and take hold of the flask without lifting it. Once the starting signal was given by the experimenter, the participants could commence filling the water from the flask in the funnel. The filling time was measured between lifting up the flask from the desk and putting it back. Once the funnel was empty, the volume of water in the glass was measured. This was done using a graduated measuring glass.[1] The participant was invited to verify the result. In all cases an agreement could be reached and the figure was filled in the protocol form. Then the experiment was set up anew and repeated. All funnel and glass experiments were videotaped for the purpose of documentation and more precise analysis.

RESULTS FROM THE EXPERIMENT

Data Preparation

The five tasks of the UoA inventory included, all in all, 14 subtasks. Subtasks were assessed on a right (1) or wrong (0) basis. The UoA overall score (UoA_O) was determined as the percentage of correct answers to the subtasks. A score for each of the five UoA tasks—RWT, BB, BD, MC, BT—was defined similarly as percentage of correct answers to the specific task's subtasks.

Performance measures in the funnel and glass task included volume variance (FE_VV), absolute volume variance (FE_VV+), filling time (FE_FT), and cash-out (FE_CO). The volume variance was calculated as actual volume minus goal volume—both measured in millilitres. Positive values for the volume variance therefore mean overshoot (FE_VV_OS) and negative values translate into undershoot (FE_VV_US). The filling time was both measured manually with a stop watch and subsequent to the exercise determined by analyzing the video shot taken. Because of technical difficulties with the

[1] More precisely, the measurement is taken from the bottom of the meniscus, which is the curved surface of the liquid. The meniscus forms because water molecules are more attracted to glass than to each other.

video camera, six participants were not recorded. In addition to this, several video recordings could not be analyzed because participants obscured the view with their bodies. If the video recording could be used to determine the filling time with high precision this result was used instead of the manually stopped time span (which was used otherwise).

As a first step in the statistical analysis, all data were carefully screened following the guidelines provided by Tabachnick & Fidell (2007). Distributions of all variables were checked for outliers and violations of the normality assumption. As in this process several skewed distributions and a few outliers were detected, transformations were applied as follows (note that a 1 or a 2 is used to refer to the first and the second run of the funnel and glass task):
FE_VV_1_SQRT=SQRT(FE_VV_l + 41)
FE_FT_1_LN = LN(FE_FT_1)
FE_FT_2_LN = LN(FE_FT_2)

For the UoA sub-task scores UoA_RWT, UoA_BB, UoA_BD, UoA_MC, and UoA_BT no transformations were found that could correct the violation of the normality assumption. Fortunately, no outliers could be detected for these scores. Re-running the procedure of outlier detection for all transformed variables identified two problematic cases that were deleted from the data set for further analysis (resulting in N = 69). Based on Mahalanobis distance no multivariate outliers could be identified in this reduced data set. To facilitate interpretation of our findings we do not use the transformed variables in the statistical analyses. Instead, we rely on non-parametric statistics in addition to parametric tests to increase the robustness of our analysis against violations of the normality assumption.

Statistical Results Regarding Performance in the Tangible Stock/Flow Test

We assess our first proposition that participants perform well by regarding (i) the volume variance, (ii) the filling time, and (iii) the aggregated performance in the tangible stock/flow experiment. The results for the two experimental runs are analyzed and presented separately.

In the first experimental run, the volume variance ranged from -40 ml to 82 ml. Average volume variance was 20.5 ml, which is significantly different from zero, $t(68)=10.09$, $p<0.001$, and is also significantly different from 10 by allowing for a ±10% tolerance zone, $t(68)=5.16$, $p<0.001$. Volume variance median is 16 ml. According to the one sample Wilcoxon signed rank test this is significantly different from both a median of zero ($Z=7.06$, $p<0.001$) and a median of 10 ($Z= 4.50$, $p<0.001$). Average overshoot is 29.3 ml (N=31), and average undershoot is –14.5 ml (N=35). Both overshoot volume variance and undershoot volume variance are significantly different from 10 and –10 respectively. The results for the second experimental run are compiled in Table 1 and Table 2.

		T-test value = 0			T-test value = 10 or -10		
	Mean	t	df	p	t	df	p
FE_VV+_2	17.41	9.377	68	.000	3.990	68	.000
FE_VV_OS_2	24.62	8.565	33	.000	5.086	33	.000
FE_VV_US_2	-12.13	-6.786	29	.000	-1.193	29	.242

Table 1: T-test results for second experimental run in the funnel and glass task

			Test value = 0		Test value = 10 or -10	
	N	Median	Z	p	Z	p
FE_VV+_2	69	14.00	6.957	.000	3.311	.001
FE_VV_OS_2	34	22.00	5.089	.000	4.157	.000
FE_VV_US_2	30	-10.50	-4.785	.000	-.878	.380

Table 2: Wilcoxon signed rank Test results for second experimental run in the funnel and glass task

These results show that proposition P1 has to be considered not-supported based on the performance measure volume variance. Participants are not able to reach the target water level within a range of ±10 % when filling water through a funnel in a glass. They either significantly overshoot or undershoot the target with one exception: undershoot in the second run does not significantly drop out the ±10 % range.

Assessing the first proposition by focusing on the filling time is more difficult. We focus on descriptive statistics first. Mean filling time is 8.09 seconds in the first run and 6.94 seconds in the second run. Minimum values are 2.44 and 2.68 seconds. This is rather fast when one considers that it takes about two seconds to fill the exact amount of water in the funnel without spouting out or flowing over the brim of the funnel. Based on the small test sample used for calibrating the cash-out function, a filling time of five seconds was set as reference value (Figure 3). Table 3 compares the participants' filling time to this reference values using both the one sample T-test and a Wilcoxon signed rank test. Both runs show a significant deviation from the reference filling time. However, in the second run participants come closer to the reference time.

				T-test value = 5			Wilcoxon Signed Rank Test value = 0	
	N	Mean	Median	t	df	p	Z	p
FE_FT_1	69	8.09	6.40	5.155	68	.000	4.326	.000
FE_FT_2	69	6.94	5.36	3.614	68	.001	2.239	.022

Table 3: Filling time test results

Thus, we conclude that we did not find support for the first proposition which stated that participants perform well in a tangible stock/flow task. The tangibility of the task apparently did not help in achieving a good performance. We also find that no learning occurs between run one and run two regarding the volume variance performance measures. A paired sample T-test shows non-significant differences in the scores for FE_VV+_1 (Mean=20.48, SD=16.87) and FE_VV+_2 (Mean=17.41, SD=15.42); t(68)=1.037, p=0.303. The non-parametric related samples sign test and the Wilcoxon signed rank tests corroborate this finding. Some learning can be observed regarding the filling time measures. The difference in the filling times FE_FT_1 (Mean=8.09,

SD=4.98) and FE_FT_2 (Mean=6.94, SD=4.46) is significant in the paired sample T-test; t(68)=2.908, p=0.005.[2] However, the progress is rather slow.

One could suspect that some sort of false learning might have occurred, specifically, participants who experienced overshoot in the first run might overreact and walk into the undershoot trap and vice versa. This, however, is not the case as Table 4 illustrates. Undershoot in the first run (FE_VV_US_1) is not significantly correlated with overshoot in the second run (FE_VV_OS_2) and vice versa. The only significant correlation between FE_VV_US_1 and FE_VV_US_2 indicates that undershoot is a rather persistent phenomenon.

			FE_VV_OS_1	FE_VV_US_1	FE_VV_OS_2	FE_VV_US_2
Spearman's rho	FE_VV_OS_1	Correlation Coefficient	1.000	.	-.376	-.138
		Sig. (2-tailed)	.	.	.185	.610
		N	31	0	14	16
	FE_VV_US_1	Correlation Coefficient	.	1.000	.037	.652*
		Sig. (2-tailed)	.	.	.888	.012
		N	0	35	17	14
	FE_VV_OS_2	Correlation Coefficient	-.376	.037	1.000	.
		Sig. (2-tailed)	.185	.888	.	.
		N	14	17	34	0
	FE_VV_US_2	Correlation Coefficient	-.138	.652*	.	1.000
		Sig. (2-tailed)	.610	.012	.	.
		N	16	14	0	30

*. Correlation is significant at the 0.05 level (2-tailed).

Table 4: Spearmen rank correlation analysis for overshoot and undershoot volume variance

Regarding the correlations between the volume variance and the filling time measures shown in Table 5 one can conclude that short filling times result in high volume variances and vice versa in the first run. However, the correlation is only weak (rho=-.208) and marginally significant (p=0.1). Interestingly, the filling times in run one and two are highly correlated, indicating that the behaviour pattern does rarely change: a participant who is rather slow in the first run will probably be also slow in the second run (perhaps with some slight improvement due to learning effects).

[2] Both non-parametric related sample tests—the sign test and the Wilcoxon signed rank tests—corroborate again this finding.

			FE_W+_1	FE_W+_2	FE_FT_1	FE_FT_2
Spearman's rho	FE_W+_1	Correlation Coefficient	1.000	-.060	-.208	-.181
		Sig. (2-tailed)	.	.623	.087	.136
		N	69	69	69	69
	FE_W+_2	Correlation Coefficient	-.060	1.000	-.006	-.085
		Sig. (2-tailed)	.623	.	.964	.487
		N	69	69	69	69
	FE_FT_1	Correlation Coefficient	-.208	-.006	1.000	.728**
		Sig. (2-tailed)	.087	.964	.	.000
		N	69	69	69	69
	FE_FT_2	Correlation Coefficient	-.181	-.085	.728**	1.000
		Sig. (2-tailed)	.136	.487	.000	.
		N	69	69	69	69

**. Correlation is significant at the 0.01 level (2-tailed).

Table 5: Spearmen rank correlation analysis for volume variance filling time measures

Summarizing this first set of analyses, we doubt whether a naturalistic decision making setting really helps making better decisions when it comes to stock and flow processes. Achievements of participants differed significantly from the objectives, even when we include a tolerance level of correct goal achievement. We do not find support for our first proposition. Next, we evaluate performance in the tangible test compared to performance in a computer-based and a paper-pencil stock/flow experiment.

Comparison of Computer Simulated and Tangible Stock/Flow Experiment

Because of the structural similarity of this study's tangible stock and flow task to the alcohol simulator task conducted by Moxnes & Jensen (2009) as a computer simulation experiment, a comparison of results seems a worthwhile endeavour. By such a comparison we can directly address our proposition P2a that postulates that participants perform better in a tangible stock/flow experiment than in a similar simulator based experiment. Moxnes & Jensen's (2009) alcohol simulator has a simple user interface showing information on the target alcohol concentration in blood, the number of bottles of beer drunken so far and the current alcohol concentration in a tabular manner. A decision on the number of bottles of beer to be drunk in the next 15 minutes is required and has to be entered. Then a button has to be pushed to advance 15 minutes. Comparing such a frugal user interface and the computer based beer drinking process to real stock and flow decision making tasks strengthens the main argument of naturalistic decision making research—that is, that apparent failures of humans to deal adequately with dynamic complexity do not result from erroneous thinking but are artefacts of the experimental method employed.

The tangible task used in this study of filling water through a funnel in a glass is the exact opposite of a computer simulation experiment. An everyday activity—filling water in a glass—is slightly complicated by introducing a second stock in the system: the water in the funnel. By this, the structural equivalence between this study's task and the alcohol drinking task is achieved. Everybody knows how funnels work. The glass funnel employed and the water coloured red ensure that the stock of liquid in the funnel is clearly visible. The target level of water in the beaker is clearly indicated using an

arrow pointing to a measuring line. Nevertheless, as we have shown in the previous section, participants perform rather poorly in this task.

Comparison of our results to Moxnes & Jensen's findings is restricted to volume variance and measures of overshoot; simulated drinking time and real filling time are not comparable. Moxnes & Jensen report for the short stomach delay time of 4.5 minutes an average blood alcohol concentration of 21% over target. In the funnel and glass task with an even shorter delay of only a few seconds caused by the funnel, average percentage volume variance is 20.90% in the first run and 17.76% in the second run. None of these two results differ significantly (on the .05 level) from the 21% in a one sample T-test; t(68)=-0,050, p=0.960, and, t(68)=-1,710, p=0.092. However, based on the Wilcoxon signed rank test the percentage deviation in the second run is significantly different from a median of 21% (Z=-2.509, p=0.012). Nevertheless, this is the only significant difference that we found and we assume it to be related to a weak learning effect in the tangible experiment while there was no second try in the computer-based task.

In summary, proposition 2a cannot be supported: on average, participants in a tangible stock/flow experiment perform similarly to participants in a computer-based task. Even the size of goal variance is similar. We continue testing the tangible experiment's outcomes with results in a paper/pencil test.

Comparison of Paper/Pencil and Tangible Stock/Flow Experiment

For reasons of completeness, we start with a description of the results for the UoA test. Figure 4 provides a histogram for the results of the UoA test. With a mean value of .560, a standard deviation of .231, a minimum performance of .0714 and a maximum of .9286, the participants' understanding of accumulation has to be rated as below appropriate.

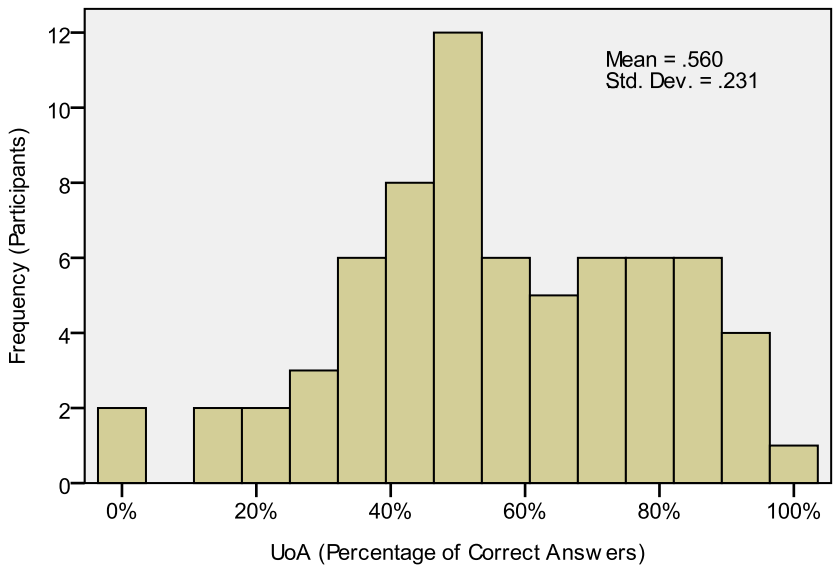


Figure 4: Histogram of the participant's UoA score

Table 6 provides an overview of participants' sub-task results based on the percentage of correctly answered items. It also shows results from other studies. We did not follow Booth Sweeney & Sterman's (2000) detailed rating approach, we rather assessed each subtask on a right or wrong basis. Therefore, a direct comparison of results is not always possible. We nevertheless included the result of the worst sub-item as approximation.

		UoA_RW	UoA_BB	UoA_BD	UoA_MC	UoA_BT
This study	N	69	69	69	69	69
	Mean	0.691	0.536	0.681	0.198	0.696
	SD	0.354	0.464	0.335	0.304	0.464
Booth Sweeney & Sterman (2000)	N				109	95
	Mean				0.1[d]	0.68[d]
Cronin et al. (2009)	N		173			
	Mean		0.376			
Osssimitz (2002)	N			154		154
	Mean			0.322		0.26
Strohhecker (2009)	N	26	26			
	Mean	0.192	0.192			

[d] Detailed coding used. Not completely comparable.

Table 6: Descriptive statistics of Understanding of Accumulation test/sub-tests

Although this study's participants achieve results in paper and pencil UoA tasks that are mostly better than the performance found in other research, understanding of accumulation is still far from optimal. All in all, this study adds to the pool of results found by previous work (Booth Sweeney & Sterman, 2000; Cronin & Gonzalez, 2007; Cronin *et al.*, 2009; Kainz & Ossimitz, 2002; Ossimitz, 2002; Sterman, 2002; Sterman & Booth Sweeney, 2002, 2007; Strohhecker, 2009). Once more, it demonstrates a fundamental 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—at least, when an "abstract" paper/pencil task is considered.

Our proposition P2b suggests that participants are able to perform better in a tangible stock/flow experiment than in paper and pencil stock/flow tasks. As the performance criteria for the two tasks are obviously not identical, direct comparison is difficult. Nevertheless, we try to come closer to an assessment of P2b by using the percentage deviation from the target value in both tasks. UoA percentage deviation is determined by subtracting the UoA percentage score of correct answers as shown in Figure 4 from a target value of 100%. Percentage deviation for the volume variance is calculated by dividing the absolute volume variance by the target volume (98 ml). Regarding the filling time a percentage deviation from the reference value of five seconds is determined analogously. Table 7 compiles the results. It shows that the percentage volume deviation from target is much lower for both runs of the funnel and glass tasks than for the paper and pencil test. While the filling time deviation is similar for the second run, for the first run a much higher mean has to be noted. Standard deviation for the filling time measures is more than five times higher than for the volume variance deviations.

	Percentage Deviation from Target				
	UoA	FE_VV+_1	FE_VV+_2	FE_FT_1	FE_FT_2
Mean	43.996	20.896	17.761	61.829	38.762
Std. Error of Mean	2.777	2.072	1.894	11.994	10.727
Median	50.000	16.327	14.286	28.000	7.200
SD	23.068	17.210	15.734	99.627	89.104
Skewness	21.649	1.412	1.249	148.610	169.591
Std. Error of Skewness	28.874	0.289	0.289	28.874	28.874
Kurtosis	-34.171	2.453	1.467	241.813	271.991
Std. Error of Kurtosis	57.010	0.570	0.570	57.010	57.010
Range	100.00	83.67	70.41	506.40	414.40
Minimum	.00	.00	.00	-51.20	-46.40
Maximum	100.00	83.67	70.41	455.20	368.00

Table 7: Descriptive statistics for percentage deviations from target in % values in both the paper-pencil and the funnel and glass tasks

Aggregation of the two performance components volume variance and filling time to an overall funnel and glass performance measure can be done using either the sum of the percentage deviations shown in Table 7 or the percentage shortfall of the actual cash-out compared to the maximum of 5 €. Table 8 shows the descriptive statistics for these two measures and the two runs. In all four cases the mean deviation from the target is larger than in the paper/pencil test. Paired sample T-tests and non-parametric Wilcoxon signed rank tests were conducted to investigate if these differences are statistically significant. Table 9 compiles the results. The only non-significant difference between paper-pencil and tangible stock/flow performance occurs in the second experimental run when the aggregate percentage deviation from target measure (FE_FT+VV+_2 = FE_PDT_2) is considered. In all other combinations participants perform better in the paper and pencil test. Proposition P2b is therefore not supported.

	Percentage Deviation from Target				
	UoA	FE_FT_+ VV+_1	FE_FT_+ VV+_2	FE_ Cashout_1	FE_ Cashout_2
Mean	0.440	0.827	0.565	0.698	0.606
Std. Error of Mean	0.028	0.117	0.107	0.038	0.043
Median	0.500	0.492	0.315	0.760	0.680
Std. Deviation	0.231	0.976	0.887	0.316	0.356
Skewness	0.216	1.355	1.495	-0.648	-0.418
Std. Error of Skewness	0.289	0.289	0.289	0.289	0.289
Kurtosis	-0.342	1.921	2.029	-0.847	-1.182
Std. Error of Kurtosis	0.570	0.570	0.570	0.570	0.570
Range	1.00	4.82	4.11	1.04	1.08
Minimum	.00	-.25	-.37	-.04	-.08
Maximum	1.00	4.56	3.74	1.00	1.00

Table 8: Descriptive statistics for aggregated percentage deviations from target in % values in both the paper-pencil and the funnel and glass tasks

	Paired sample T-test					Related sample Wilcoxon Signed Rank Test	
	Paired Diff.						
UoA_PDT paired with	Mean	SD	t	df	p	Z	p
FE_PDT_1	-0.387	0.955	-3.370	68	0.001	2.198	0.028
FE_PDT_2	-0.125	0.907	-1.147	68	0.255	-0.963	0.851
FE_Cashout_PDT_1	-0.258	0.368	-5.818	68	0.000	4.729	0.000
FE_Cashout_PDT_2	-0.166	0.437	-3.150	68	0.002	1.445	0.005

Table 9: Paired sample sample T-test and Wilcoxon Signed rank test results

Additionally, the null hypothesis was tested—using Friedman's two-way analysis of variance by ranks—that the distributions of each of the pairs shown in Table 9 are the same. This hypothesis could only be rejected for UoA_PDT and FE_Cashout_PDT_1, $\chi^2(1)=23.529$, $p<0.001$. For all other pairs the distributions have to be regarded as equal. This finding corroborates the rejection of P2b.

Investigation of the Relation between Paper-Pencil and Tangible Stock/Flow Performance

With proposition P3 we want to investigate if there is a relation between performance in naturalistic and in abstract dynamic decision making. Therefore, P3 suggests that participants performing well in the tangible stock/flow task will also show a good understanding of accumulation in a paper-pencil test and vice versa.

For testing this proposition we use first non-parametric and parametric bivariate correlation analysis and second a quasi-experimental approach that is evaluated by both T-test and Wilcoxon signed rank test. For P3 to hold, positive correlations between performance in paper and pencil test and funnel and glass task have to be found. No or significantly negative correlation between funnel and glass task performance and UoA would mean that P3 is not supported.

We calculate Spearman's rank correlation coefficient (and also Pearson correlation) for UoA (better performance means higher percentages), absolute volume variance (better performance means lower FE_VV+) and filling time (better performance means shorter times FE_FT). As can be seen from Table 10, only little or no associations can be found; all correlations are non-significant with the exception of a marginal significant Pearson correlation coefficient between filling time in the first run and UoA. Based on this finding, we have to consider P3 as not supported.

		UoA	FE_VV+_1	FE_FT_1	FE_VV+_2	FE_FT_2
UoA	Spearman	1.000	.016	-.142	.119	-.068
	p	-	.897	.244	.330	.581
	Pearson	1.000	.035	-.210	.174	-.071
	p	-	.778	.084*	.152	.560

*, **, *** indicates significance at the 90%, 95%, and 99% level, respectively.

Table 10: Spearman and Pearson bivariate correlations (N=69)

Rejection of P3 is corroborated when analyzing UoA sub-task results. As with the aggregate UoA score, we also do not find significant correlations between FE results and sub-task results. There is one exception. In the second experimental run, volume

variance is weakly associated with performance in the budget deficit paper and pencil task. However, the correlation is positive, meaning, that a better understanding of the budget deficit problem results in larger volume variance, which also contradicts P3.

Spearman		FE_VV+_1	FE_FT_1	FE_VV+_2	FE_FT_2
UoA_RWT	Correlation	.020	-.177	-.032	-.017
	p	.873	.145	.793	.888
UoA_BB	Correlation	.044	-.103	-.015	-.003
	p	.721	.400	.903	.983
UoA_BD	Correlation	-.056	-.049	.242	-.099
	p	.648	.686	.045**	.416
UoA_MC	Correlation	.044	-.034	-.002	.085
	p	.721	.783	.987	.490
UoA_BT	Correlation	-.087	-.137	.177	-.130
	p	.477	.262	.147	.288

Pearson		FE_VV+_1	FE_FT_1	FE_VV+_2	FE_FT_2
UoA_RWT	Correlation	.048	-.278	.052	-.097
	p	.694	.021**	.671	.426
UoA_BB	Correlation	.079	-.097	-.002	-.107
	p	.521	.427	.986	.381
UoA_BD	Correlation	-.053	-.129	.231	-.087
	p	.667	.289	.056*	.479
UoA_MC	Correlation	.134	-.040	.074	.166
	p	.272	.746	.548	.173
UoA_BT	Correlation	-.100	-.085	.120	-.075
	p	.415	.490	.324	.543

*, **, *** indicates significance at the 90%, 95%, and 99% level, respectively.

Table 11: Spearman and Pearson bivariate correlations for UoA sub-task and funnel and glass performance (N=69)

This finding holds also when the aggregate percentage deviations from target performance measures as introduced in the previous section are used instead of filling time and volume variance.

In a final quasi-experimental analysis we assign the cases to two groups based on the performance results achieved in the funnel and glass task. For instance, participants that have achieved a volume variance in the first run (FE_VV+_1) of less or equal to 10 are classed together and all participants that have performed worse are put in the second group. This procedure is repeated for the measures, FE_VV+_2, FE_FT_1, and FE_FT_2. In the cases of filling time the cut-off point was set to 5 seconds. Independent T-tests and Mann Whitney U tests were run testing for differences between the groups regarding both the aggregate UoA measure and all five UoA sub-task measures. Not once was the null hypothesis significantly rejected that distributions are the same across the two groups. Therefore, refutation of P3 is corroborated.

Not finding support for P3 is an interesting result since performance over the complete sample of participants is low. However, it seems that not necessarily somebody achieving a relatively high score in the UoA test also does well in the tangible test; a participant with a low score on the UoA test could well achieve a high performance in the tangible part of the experiment. The two tests seem to require different cognitive

abilities: although from an abstract structural point of view the two tasks are similar (even nearly identical), the difference in their tangibility results in different individual performance.

CONCLUSIONS, LIMITATIONS, AND FURTHER RESEARCH

We used a combination of a tangible stock/flow task with a paper-based understanding of accumulation test to assess the validity of three propositions in a laboratory experiment. First, we tested whether participants in the tangible test perform well in absolute terms. Second, we compared their performance to a benchmark. Third, we try to find correlations between the performance in the UoA test and the tangible test. We did not find support for any of the three propositions that we stated. As for the classical studies in dynamic decision making, also with a tangible task participants do not achieve good results. Compared to a paper-pencil or a computer-based understanding of accumulation task they do not manage to perform better in the tangible set-up. Furthermore, performance in the tangible task cannot be correlated to performance in the paper/pencil UoA test.

Proponents of naturalistic decision making claim that people perform better in natural situations (compared to the results in laboratory experiments of dynamic decision making). Some researchers of dynamic decision making support this but argue that today's complex decisions resemble more the abstract task of the laboratory than the daily life examples of naturalistic decision making. However, our study suggests that the assumption of increased performance in natural decision making situations does not hold for stock/flow tasks with a delay. We see two possible explanations for this finding: first, people have not acquired heuristics for this kind of tasks as they have for other tasks, which let them perform well in naturalistic settings. There is no heuristic "that makes us smart" regarding accumulation processes involving a delay. Second, although we used a tangible task for participants to fulfil, we still conducted the experiment in a laboratory setting. Thus, benefits of naturalistic decision situations could maybe not be achieved in this artificial situation.

In addition, the results regarding the third proposition (no correlation between tangible and paper-pencil test) render it doubtful whether performance in the two types of tasks should be related at all. Again, we offer two explanations for this result. First, our study suggests that people use different cognitive capabilities and that knowledge (or heuristics) used in a tangible stock/flow task cannot be transferred to a paper-pencil task and vice versa. In this sense, even if people were to achieve good performance in naturalistic tasks that would not guarantee that they also do well in abstract tasks and the other way around. Thus, as long as both types of tasks exist and are relevant decision makings situations, they need to be considered separately. Second, although the tasks used in both versions (tangible and paper-pencil) are similar in structural terms, they are not absolutely identical in terms of the framing of the task. For example, the sub-task from the UoA test that comes closest to the tangible task (the manufacturing case MC) uses the context of a production company as compared to the rather down-to-earth filling of a glass of water in the tangible test.

In summary, our findings corroborate Funke's (2001) and Bakken's (2008) conception that indeed dynamic decision making and naturalistic decision making have more in common than that they are different. The two methodologically-based explanations of our findings suggest natural ways for follow up research. Although difficult, striving for a more every-day situation could take away the criticism that we actually did not put people in a real naturalistic decision making setting. The methodological problems related to this, however, are substantial: how can you make people do what you want to do outside the laboratory, why would they fill-out a questionnaire there, would you be able to measure exactly their performance, are just some of the questions that would need to be answered. The second methodological issue could be tackled more easily by better aligning the two tests. For instance, in the paper-pencil UoA test a question could be included on how much water to fill-in (obviously, the target volume) and how long one would need to wait for the water to fill the glass.

REFERENCES

Bakken BE. 2008. On improving dynamic decision-making: implications from multiple-process cognitive theory. *Systems Research and Behavioral Science* **25**(4): 493-501

Booth Sweeney L, Sterman JD. 2000. Bathtub dynamics: initial results of a systems thinking inventory. *System Dynamics Review* **16**(4): 249-294

Brehmer B. 1992. Dynamic decision making: human control of complex systems. *Acta Psychologica* **81**: 211-241

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

Cronin MA, Gonzalez C, Sterman JD. 2009. Why don't well-educated adults understand accumulation? A challenge to researchers, educators, and citizens. *Organizational Behavior & Human Decision Processes* **108**(1): 116-130

Dörner D. 1980. On the difficulties people have in dealing with complexity. *Simulations and Games* **11**(1): 87-106

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

Dörner D, Kreuzig HW, Reither F. 1994. *Lohhausen. Vom Umgang mit Unbestimmtheit und Komplexität*. Huber: Bern et al.

Edwards W. 1962. Dynamic decision theory and probabilistic information processing. *Human Factors* **4**(2): 59-73

Forrester JW. 1961. *Industrial dynamics*. Productivity Press: Cambridge, MA

Friedman D, Cassar A, Selten R. 2004. *Economics lab: An intensive course in experimental economics*. Routledge: London

Funke J. 2001. Dynamic systems as tools for analysing human judgement. *Thinking and Reasoning* **7**(1): 69-89

Guala F. 2005. *The methodology of experimental economics*. Cambridge University Press: Cambridge

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

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. In SD Society (Ed.), *20th International Conference of the System Dynamics Society*: Palermo

Klein G. 2008. Naturalistic decision making. *Human Factors: The Journal of the Human Factors and Ergonomics Society* **50**(3): 456-460

Lipshitz R, Klein G, Carroll JS. 2006. Introduction to the special issue. Naturalistic decision making and organizational decision making: Exploring the intersections. *Organization Studies* **27**(7): 917-923

Lipshitz R, Klein G, Orasanu J, Salas E. 2001. Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making* **14**(5): 331-352

Moxnes E. 1998. Not only the tragedy of the commons: Misperceptions of bioeconomics. *Management Science* **44**(9): 1234-1248

Moxnes E. 2011. A unifying theory of local and global overshoot−From discounting to simulation, *5th European System Dynamics Workshop*: Frankfurt am Main

Moxnes E, Jensen L. 2009. Drunker than intended: Misperceptions and information treatments. *Drug and Alcohol Dependence* **105**(1–2): 63-70

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

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

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

Sengupta K, Abdel-Hamid TK. 1993. Alternative conceptions of feedback in dynamic decision environments: An experimental investigation. *Management Science* **39**(4): 411-428

Smith VL. 1976. Experimental economics: Induced value theory. *American Economic Review* **66**(2): 274

Smith VL. 1982. Microeconomic system as an experimental science. *American Economic Review* **72**(5): 923

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

Sterman JD. 1994. Learning in and about complex systems. *System Dynamics Review* **10**(2-3): 291–330

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

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

Sterman JD, 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

Strohhecker J. 2009. A pilot study for testing the effect of stock and flow thinking on stock and flow management performance. In J Strohhecker, A Größler (Eds.), *Strategisches und operatives Produktionsmanagement - Empirie und Simulation*: 285-305. Gabler: Wiesbaden

Tabachnick BG, Fidell LS. 2007. *Using multivariate statistics* (5 ed.). Pearson/A&B: Boston, MA, London

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

Zsambok CE, Klein GA. 1997. *Naturalistic decision making*. L. Erlbaum Associates: Mahwah, N.J.