The Attempt to Boost Stock Flow Performance by Using a General Problem Solving Strategy and a Reappearing Gender Effect

Vivien Röder

Technische Universität Chemnitz
W.-Raabe-Str. 43, 09120 Chemnitz, Germany,
+49-371-531 34143
vivien.roeder@psychologie.tu-chemnitz.de

Abstract

It was over 16 years ago that Booth Sweeney and Sterman (2000) first wrote about the poor performance of well educated adults on simple dynamic systems. Boosting performance in those stock flow tasks has been a research topic ever since but still lacks a resounding success. A recent study looked at the impact of using a general compared to a conventional problem solving strategy when confronted with mathematical and economic problems (Youssef-Shallala, Ayres, Schubert & Sweller, 2014). This study gave the idea to adapt this manipulation and to apply it to stock flow problems. In a think aloud setting participants solved four drawing stock flow tasks and one that was followed by four questions, using either a conventional (CPS) or a general problem solving (GPS) strategy. No significant improvement could be found considering the used strategy alone. However, the average solution rate for the four drawing problems reached almost 70% which is far higher than in previous studies. Men even averaged on over 90%, thus probably reaching a ceiling effect. Women performed significantly worse but seemed to profit from the GPS strategy. Overall solution rates dropped in the question task. No strategy advantage was found but the gender effect remained.

Keywords: dynamic systems; stock–flow systems; general problem solving; problem solving strategy; think aloud; gender effect

Introduction

Dealing with dynamic systems is a part of our everyday lives. On a personal level we have to manage our bank account, ponder whether or not the fuel in the car will be sufficient to reach the destination in case of a traffic jam during a snow storm, or making sure that there is always enough food in the refrigerator for the family. On a bigger scale dealing with complex dynamic systems might involve decisions concerning the GDP or global warming and therefore affects millions of people. Sometimes those systems are more, sometimes a little less complex but as we are confronted with them daily, it is desirable that everybody does understand them well enough to make founded decision.

However, even stock flow tasks, a very simple form of dynamic problems, have been shown to lack understanding even among highly educated MIT students (Booth Sweeney & Sterman, 2000). Since the often cited stock flow study from Booth Sweeney and Sterman (2000) many more followed, manipulating different aspects of the task setting to find a way to boost the repeatedly found poor performance (e. g. Armenia, Onori, & Bertini, 2004; Brockhaus, Arnold, Schwarz, & Sedlmeier, 2013; Brunstein, Gonzalez & Kanter, 2010; Cronin, Gonzalez, & Sterman, 2009; Kainz & Ossimitz, 2002; Kapmeier, 2004; Korzilius, Raaijmakers, Rouwette and Vennix, 2014; Schwarz, Epperlein, Brockhaus, & Sedlmeier, 2013). The examined variables included domain related knowledge, representation format, task context, participants' motivation, static vs. dynamic representation or the cognitive capacity. The outcome of all those studies yielded rather ambiguous results, however

Other domains also experience the problem of low solution rates for tasks and search for options to improve them. A recent study by Youssef-Shalala and colleagues (2014) worked with junior and senior high school students. In four independent experiments, they tried to boost solution rates for mathematic or economic problems using a general (GPS) compared to a conventional problem solving (CPS) strategy or compared to worked examples. The idea was based on the cognitive load theory (Sweller, Ayres & Kalyuga, 2011) In the GPS strategy condition participants were encouraged to first use a goal free strategy, so that they could randomly generate and test ideas towards unknown problems. The test problems at the end were either similar to the acquisition problems or required a far transfer. Students differed in their prior knowledge about the topics. On many test problems students who had used a GPS strategy performed better than in the CPS or the worked example condition. The authors also reported an interaction between the strategy and ability of the students, with lower ability students benefitting more from the GPS strategy than students with high abilities. Overall Yousef-Shalala et al. concluded that applying a GPS strategy might be beneficial especially for low ability students. Based upon these findings, the present study adapted this manipulation and applied it to stock flow problems

As shown above, manipulations with the goal to boost performance in tasks about simple dynamic systems only yielded ambiguous results in the past. One reappearing side effect that seemed to be independent of other task-specific or participant-specific factors (Schwarz, 2016) was a repeatedly found gender effect with men outperforming women on average. Although not every study reported the performance separately for men and women, multiple researches have at least mentioned a gender effect (Booth Sweeney & Sterman, 2000; Brockhaus et al., 2013; Jensen, 2002; Kainz & Ossimitz, 2002; Kapmeier, 2004; Kasperidus, Langfelder, & Biber, 2006; Schwarz et al., 2013; Veldhuis & Korzilius, 2012) and a recent study specifically looked at the gender difference in stock flow tasks (Röder, Sedlmeier and Schwarz, under review). Therefore, it was anticipated that this study would also display a gender effect.

Another aim of this study was to broaden the data base for stock flow problems solved in a think aloud setting. The think aloud method is characterized by speaking out loud whatever comes to mind while working on a certain task. The goal is to get an insight into the cognition at work and to identify reasoning patterns. Up to date only one big think aloud study in the context of stock flow tasks is known to the author. Korzilius, Raaijmakers, Rouwette and Vennix (2014) asked 50 participants to solve a stock flow problem while thinking aloud. It was a stock flow problem followed by four questions – a task format that had been used before multiple times and is commonly referred to as department store task or discontinuous task (Cronin et al., 2009; Schwarz et al., 2013; Sterman, 2002). The researchers found no performance difference for this type of stock flow task between the think aloud condition and a written control group with 65 participants. Results were also comparable with prior experiments using a discontinuous stock flow tasks. Although the task with the four questions is common in stock flow research, much of the stock flow research conducted also used an operationalization in which participants had to draw the stock over time based on the in- and outflow. Those tasks are commonly framed in the context of a bath tub and therefore are often referred to as bathtub task (Booth Sweeney & Sterman, 2000; Kapmeier, 2014; Ossimitz, 2002; Schwarz et al., 2013). Korzilius and colleagues did not include a drawing task in their experiment, however. This study now closes this gap by using four drawing tasks. One discontinuous stock flow task with four questions was also included for comparability reasons. The qualitative analysis of the think aloud protocols shall not be reported in this paper here as the focus is on the solution rates itself.

The central questions addressed in this study were whether a modification of the used problem solving strategy (GPS vs. CPS) has an impact on the stock flow solution rates. The hypothesis was that a GPS strategy yields better results than a CPS strategy. Furthermore, the impact of gender was of interest. Following past research a better performance for men was anticipated. In addition the study wanted to broaden the data base for stock flow tasks solved in a think aloud setting. Finally, the author was interested in possible moderators such as the grade in mathematics, the time used to solve the stock flow tasks or the motivation.

Method

Participants

As the study was conceptualized as a think aloud study the sample size was rather small with a total of 20 participants. Precisely 50 percent of them were female. Participants were recruited by mailing list at the Technische Universität Chemnitz in Germany and received either course credit or a small candy bar. The mean age was 22.4 years (*SD* = 3.9) and the vast majority (90%) studied psychology. One participant had already graduated and had a job. Three remembered that they had participated in an earlier stock flow study.

Materials

Think aloud tasks. Following van Someren and colleagues (1994) who proposed a short warm up of up to 15 minutes to give the participants the option to get accustomed to the think aloud setting, each participant first received an explanation about the concept of thinking aloud and was then given two tasks to familiarize themselves with the think aloud setting. In the first task participants were asked how many windows had been in their parents flat or house (Ericsson and Simon, 1993). The second problem asked them to measure 4 liters with the help of a 3 and a 5 liter water jar. Both tasks only served to get participants at ease with the think aloud setting and were not evaluated later on.

Stock flow tasks. To test the participants' understanding of SF problems, we used five different tasks that have been commonly employed in stock flow research (e.g. Booth Sweeney & Sterman, 2000; Cronin et al., 2009; Kapmeier et al., 2014; Kasperidus et al., 2006; Sterman, 2002). The flow information was always presented in a line graph depicting the rates of in- and outflow per minute with two distinguishable (e.g., dotted and solid) lines. Following the procedure of Booth Sweeney and Sterman (2000), participants were given a short introduction explaining the general concept of simple dynamic systems before they were confronted with the problems. Stock flow tasks were presented in alternating sequences, however, the discontinuous task with the four questions was always presented last.

Four tasks were given in the context of water flowing in and out of a bath tub over a period of 16 minutes and participants were asked to draw the development of the stock in an empty diagram underneath. All drawings were later rated by two independent raters who were neither aware of the condition (GPS or CPS, male or female) nor of any other possible influence factors such as grade in mathematics or time used to solve the task. The rating followed five criteria first used by Schwarz et al. (2013) which in turn were based on the seven criteria proposed by Booth Sweeney and Sterman (2000). Each stock drawing resulted in a score between 0 and 5 points which were converted into percentages. The score of the two raters was averaged. Interrater reliability is reported in the result section. The flow patterns of the four drawing tasks and their solutions can be seen in Figure 1. According to their shapes they were named K ("Keil" as the German word for wedge), G (Gaussian distribution curve), P (parallel) and W (w-shaped inflow pattern).

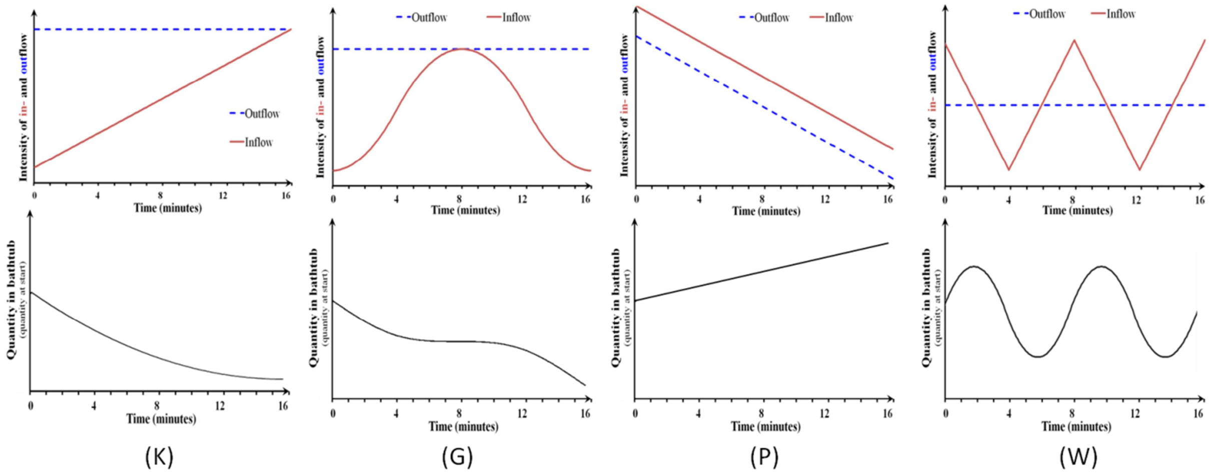
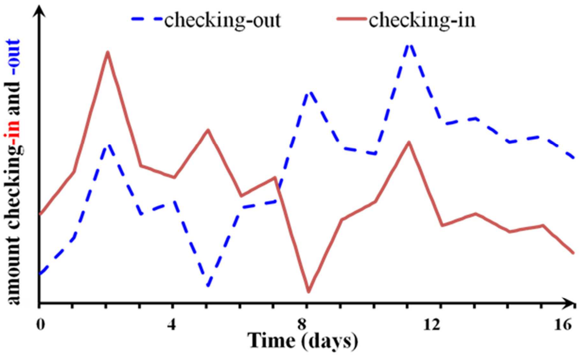


Fig. 1. The four drawing stock flow tasks used to test participants' understanding of simple dynamic systems. Previous studies have shown a variance in their difficulty. In- and outflow were represented by different colored and shaped (dotted, solid) lines to avoid confusion, especially when the lines cross. There were no precise numbers given so that the participants' general understanding of the dynamic system was tested rather than their calculation abilities. The top row shows the tasks given, the bottom row the corresponding solutions.

To compare the think aloud results in this study with the ones found in the think aloud setting by Korzilius and colleagues (2014) and to have a broader overall view on the stock flow performance I also used one task for which participants had to answer four questions instead of drawing the development of the stock over time. The context for this task were guests checking in and out of a hotel. In line with previous studies the flow pattern was more ragged than in the drawing tasks. In- and out-flow lines crossed ones. Minimum and maximum of the flows as well as the maximum net inflow and the maximum net outflow have been common mistakes in the past and were clearly distinguishable in the flow diagram (Figure 2).



Please answer the following questions:

1. During which day(s) did the most people check in to the hotel?

   Day(s) ________ O Cannot be determined

2. During which day(s) did the most people check out of the hotel?

   Day(s) ________ O Cannot be determined

3. During which day(s) were the most people in the hotel?

   Day(s) ________ O Cannot be determined

4. During which day(s) were the fewest people in the hotel?

   Day(s) ________ O Cannot be determined

Fig. 2. The discontinuous (D) hotel guest tasks with four questions used to test participants' understanding of simple dynamic systems. In- and outflow were represented by different colored and shaped (dotted, solid) lines to avoid confusion, especially when the lines cross. There were no precise numbers given so that the participants' general understanding of the dynamic system was tested rather than their calculation abilities.

In the CPS condition the tasks were handed out and participants were asked to either draw the development of the stock or answer the four questions right away. This is congruent with previous research. In the GPS condition however, the first two stock flow tasks were each preceded by a page only showing the flow diagram and information about the context. No question was asked concerning the development of the stock. Instead participants were asked to look at the diagram and to describe what they were seeing. They were only very broadly asked what information they could extract from the diagram. The aim was that they were not focused on a certain task but kept an open mind and kept thinking in all possible directions. Only after they had described what information they had found in the diagram they proceeded to the same questions about the development of the stock as in the CPS condition.

Other measures. After participants finished with the five stock flow tasks, a demographic questionnaire followed. Data on sex, age, grade in mathematics on the high school diploma, educational background, motivation, previous stock flow participation, interest in mathematics and riddles as well as how experienced they evaluated themselves in reading and drawing diagrams was collected.

Design and Procedure

A 2 x 2 factorial design was used, with the first factor being the applied problem solving strategy (CPS vs. GPS) and the second factor being the gender of the participants. The procedure was divided into four parts: After a (i) short standardized instruction explaining the general purpose and the think aloud method, participants, who were all tested in a one to one setting, were confronted with (ii) two tasks to practice to think aloud. After that each man and woman (iii) was randomly assigned to either the CPS or the GPS condition and received an introduction to the general idea of simple dynamic systems and the five stock flow tasks in alternating order followed by (iv) a demographic questionnaire. At the end, every participant could voice his or her questions to the female experimenter and was given feedback on their solution. Before leaving, participants received either course credit or a small candy bar.

Results

Stock flow drawing tasks

Stock drawings were rated by two independent raters who were unaware of the study aims. To test the interrater reliability, Cohens Kappa was calculated. It ranged from $\kappa = 1$ for the K-task, to $\kappa = .94$ for the G- and P-task, to $\kappa = .8$ for the W-task. So overall the interrater reliability was very satisfying and the average of the two ratings was used for further analysis.

In the past the four stock flow tasks used had shown to vary in difficulty (Cronin et al., 2009; Röder, Sedlmeier & Schwarz, under review). The same was the case in the present study. In this study the P-task was the easiest with an 80% (*SD* = 36.6) solution rate, followed by W (67.5%, *SD* = 32.3), K (66%, *SD* = 43.1) and G (64.5%, *SD* = 43.5). Table 1 shows the solution rates for every task in the two problem solving conditions differentiated by gender. As the pattern of the performance was the same in all tasks, the solution rates were aggregated for further analysis.

Table 1. Mean Solution Rates in Percent for the Five Stock Flow Tasks Separated for Problem Solving Strategy (Conventional Problem Solving vs. General Problem Solving Strategy) and Gender.

		K	G	P	W	Mean 4 drawing tasks	D Q1 & Q2	D Q3 & Q4
CPS	Women (n=5)	24.0	18.0	60.0	44.0	36.5	80.0	0.0
	Men (=5)	92.0	94.0	100.0	84.0	92.5	90.0	80.0
GPS	Women (n=5)	52.0	54.0	64.0	48.0	54.5	100	0.0
	Men (=5)	96.0	92.0	96.0	94.0	94.5	100	60.0

To get a first impression whether the two factors influence the stock flow solution rates, an ANOVA was conducted. Even with the small sample size a substantial effect was found, $F(3, 16) = 5.39$, $p = .009$, $\eta_p^2 = .502$. However, the manipulation on the problem solving strategy did not show a main effect $F(1, 16) = 0.66$, $p = .430$. Nonetheless, data suggests a small to medium effect size $\eta_p^2 = .039$. As expected, a main effect for gender was found, $F(1, 16) = 15.09$, $p = .001$, $\eta_p^2 = .485$. Seemingly no interaction was present, $F(1, 16) = 0.42$, $p = .527$, but once again the effect size points to a small to medium effect $\eta_p^2 = .026$. A visual analysis (Figure 3) clearly indicates that the men's solution rates seem to hit the ceiling in both the CPS and the GPS condition. For the women however, a difference between the applied problem solving strategies can be made out, favoring the GPS as anticipated.

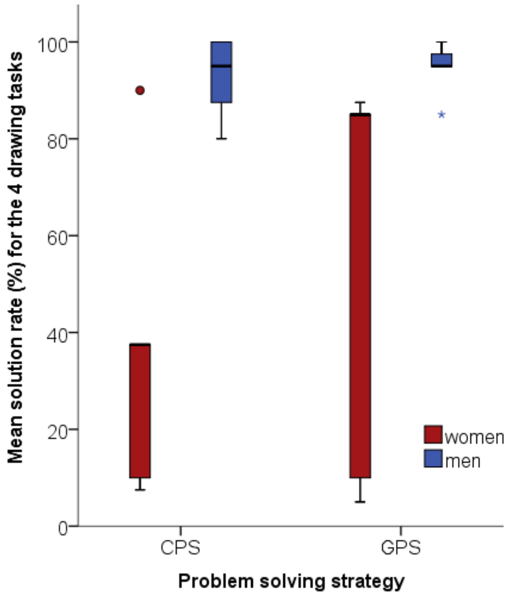


Fig. 3. The mean solution rates for the averaged performance in the four drawing tasks separated by the applied problem solving strategy and participants' gender.

3.2. Stock flow discontinuous task

For the discontinuous stock flow task a differentiation between questions 1 and 2 is not reported, as 19 out of 20 people either answered both correct or both incorrect. For questions 3 and 4 the congruency even hit 20 out of 20. Therefore questions 1 and 2 as well as 3 and 4 shall be always considered together here.

The main interest was on questions 3 and 4 (when were the most/ fewest people in the hotel) as questions 1 and 2 (which day[s] did most people check in to/ out of the hotel) are not concerned with the stock. Questions 1 and 2 rather test the general understanding and the ability to read the graph and are usually answered correctly by almost everybody. In this study 92.5% of the answers were correct for questions 1 and 2. As expected, performance dropped strongly when considering questions 3 and 4. Overall only about every third participant gave the correct answer (35%). So the percentage of correct solutions differed greatly between the drawing (69.5% overall) tasks and this discontinuous stock flow task with the questions.

Group differences were partly very strong again. No inferential statistics are reported for the factor problem solving strategy as the difference between conditions was only a single person (4 people with correct answers in the CPS, 3 in the GPS strategy condition) and could therefore easily be caused by chance. However, gender shows a shocking clear performance difference. None of the women but 7 out of 10 men gave the right answer, *t*(9) = -4.58, *p* = .001, 95% CI [-1.05, -0.35], *d* = -3.06[1]. The percentage of the correct solution rates in the different conditions for the discontinuous task can also be seen in Table 1.

3.4. Other measures

Demographic data that promised to be of explanatory help based on prior research was included in to a multiple regression analysis. As solution rates differed substantially between the four drawing tasks and the discontinuous stock flow task, regression analysis were conducted separately. The results can be found in table 2. As foreshadowed above, gender is the best predictor. Besides the motivation and the time spent to work on the stock flow tasks seem to be valid predictors, at least for the four drawing tasks. Contrary to prior research, the last school grade in mathematics did not seem to have an influence.

Table 2. Linear Regression Models with the Average Solution Rate for the Mean Four Drawing Tasks or the Correct Solution Rates for Questions 3 and 4 in the Discontinuous Stock Flow Task as

[1] Due to the fact that none of the women had a correct answer, the variances for men and women differed significantly (Levene’s test). Results reported here are adjusted accordingly.

the Dependent Variable and Six Predictors. The Model for the Drawing Tasks Can Explain 70.6% of the Variance. The Model for the Discontinuous Task Can Explain 68.3% of the Variance.

	Dependent variable: mean four drawing tasks			Dependent variable: discontinuous task		
Predictor	β	*t*	*p*	β	*t*	*p*
Intercept		-0.06	.952		-1.21	.249
Problem solving strategy	0.24	1.46	.168	-0.09	-0.54	.596
Gender	0.50	3.04	.010	0.61	3.64	.003
Grade in mathematics	0.02	0.10	.924	0.24	1.09	.297
Motivation	0.39	2.10	.056	0.23	1.13	.280
Time used to work on drawing tasks/	-0.42	-2.35	.035			
discontinuous task				-0.22	1.23	.242
Self-rated experience with graphs	0.23	1.32	.211	0.08	0.43	.672

Discussion

This study examined the performance on stock flow tasks in a think aloud setting for two different problem solving strategies (GPS vs. CPS) and separately for men and women. Overall it can be said that the solution rates were extraordinary high for four drawing tasks compared with previous research. Here a mean of 69.5 percent was achieved compared to 28-48 percent in comparable studies (Booth-Sweeney & Sterman, 2000; Kapmeier, 2004; Röder et al., under review). Although no significance was reached, the effect size suggests there might be a small effect for the factor problem solving strategy which seems to hold especially true if only the women are considered. Gender remains a big issue with regard to the performance in the stock flow tasks. For the four drawing tasks men hit a mean solution rate of over 90%. In the stock flow task with the four questions, the gender effect was even more pronounced. However, no problem solving strategy effect could be found in this final task. The overall solution rates were comparable to previous research for this task.

The main aim of this study had been to find a way to boost the repeatedly poor performance in stock flow tasks. The main manipulation used, was directing participants towards a GPS instead of a CPS. This manipulation did not yield a resounding success for all participants and is therefore in line with prior aims to boost stock flow performance (Brockhaus et al., 2013; Cronin et al., 2009; Kainz & Ossimitz, 2002; Kapmeier, 2004; Schwarz et al., 2013). However, a small success for the GPS strategy, at least in the four drawing tasks ,can be seen when only considering the women. This is also not unexpected, remembering that the women showed an overall worse performance. Yousef-Shalala et al. (2014) also reported that the GPS strategy seemed to be especially beneficial for low ability students. Women can be seen here as low ability students. The men already reached ‘too high’ solution rates in the CPS condition so that a ceiling effect might have occurred and no further improvement was possible. In the discontinuous stock flow task, no problem solving strategy effect could be observed. One reason for that might be that the GPS strategy was only forced onto the participants in the first two drawing tasks and was hoped to be used for the following three tasks as well. However, we know from other research, that transfer is seldom that easy (Reed, 1999; Sedlmeier, 1998). So, one approach for a follow-up study might be, to bring participants to use a GPS strategy right before given a discontinuous task followed by the four questions.

Although the applied problem solving strategy was not as successful as hoped for, the overall solution rate for the four drawing tasks was extremely high which is a success in itself. The question that remains: Why was the solution rate in the four drawing tasks so much higher than in past studies? Could it be that the think aloud setting has to do with that? Although the experimenter was sitting behind the participants, participants were probably more aware of her than in settings were a group of people works at the tasks simultaneously. As most students do not want to make a fool of themselves, this setting could have caused a higher motivation to do well on the tasks. Motivation was a good predictor for performance, especially in the four drawing tasks. Although Cronin et al. (2009) did not report an effect for motivation, one can argue that the operationalization for motivation in this study was not optimal. Here motivation was not manipulated but assessed by self-report.

Another good predictor for performance was the time spent on the stock flow tasks. Speaking out loud ones thoughts does require more time than working on tasks in silence. Therefore the time spent on the stock flow tasks was higher than in normal stock flow studies. This could be another explanation for the elevated solution rates in the four drawing tasks. However, Korzilius et al. (2014) did not find extremely elevated solution rates in their think aloud setting but neither did I for the discontinuous stock flow task with the four questions. So, is there a difference between drawing tasks and stock flow task for which people have to answer questions about the stock? And the even more important question: what would that difference be?

The 35% correct solutions for questions 3 and 4 are still almost twice as high as the average solution rates reported in the think aloud study by Korzilius et al. (2014) but are more in line with past stock flow research using a version of the department store task (Kapmeier, 2004; Lyneis & Lyneis, 2003; Pala & Vennix, 2005; Sterman, 2002). In an overview by Pala and Vennix (2005) it can be seen that the percentage of correct solution rates in the discontinuous stock flow task differs between studies. However, one effect that seemed to occur in all but one reported study was that the solution rate for question 3 was slightly higher than for question 4 which was not replicated here. As sample size was small and only 7 participants answered questions 3 and 4 correctly, I suggest to put not too much weight on the non-existing difference between questions 3 and 4 in the present study.

In this study the gender difference was very pronounced again and became highly significant even despite the small sample size. Taking this reappearing gender effect in mind, I strongly suggest to look at that issue more in detail. In the present study, the men did not really need any further help to solve the stock flow tasks. Women did. Past research also identified a need of men to improve their stock flow reasoning skills. So they shall not be forgotten but it is possible that men and women need different things to help them understand those simple dynamic systems. Hopefully the analysis of the think aloud transcripts will shed some light on that issue. Furthermore I consider it very important for future stock flow papers to always report results separately for men and women as results could be biased otherwise as it would have been in the present study.

Conclusion

Dynamic systems, such as the GDP, climate change, money in a bank account, guests in a hotel or the water level in a bathtub, affect people in their daily lives. An understanding of such real-life systems constitutes the basis for decisions that can affect not only the individual but also the individual’s environment and if politicians or company owners are making the decisions, millions of people can be affected. The simplest form of such a dynamic system is a stock flow system. Understanding how flows change a stock over time in such a simple system is the basic key to understanding complex dynamic systems. Boosting this understanding should be one goal for system dynamic researchers. The present study found much better solutions rates for stock flow tasks in which the solution was a drawing of the development of the stock over time than past research did. Multiple questions remain however.

One of those questions is whether the setting does influence the results. The think aloud setting used here is rather unusual. Another question is what roles motivation and processing time play. Whether or not the applied problem solving strategy boosts the understanding of simple dynamic systems cannot be said with definiteness just yet. It seems as if women do profit from a general problem solving strategy. Overall, gender remains a big influence variable with a clear advantage for the men.

References

Armenia, S., Onori, R., & Bertini, A. (2004). Bathtub Dynamics at the "Tor Vergata" University in Rome, Italy. In M. Kennedy, G. W. Winch, R. S. Langer, J. I. Rowe, & J. M. Yanni (Eds.), *The 22st International Conference of the System Dynamics Society*. Oxford, England: System Dynamics Society.

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

Brockhaus, F., Arnold, J., Schwarz, M., & Sedlmeier, P. (2013). Does the modification of the representation format affect Stock-Flow thinking? In R. Eberlein & I. J. Martínez-Moyano (Eds.), *Proceedings of the 31st International Conference of the System Dynamics Society*. Albany, NY: System Dynamics Society. Retrieved from www.systemdynamics.org/conferences/2013/proceed/papers/P1253.pdf

Brunstein, A., Gonzalez, C., & Kanter, S. (2010). Effects of domain experience in the stock – flow failure. *System Dynamics Review*, *26*(4), 347–354. http://doi.org/10.1002/sdr

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. http://doi.org/10.1016/j.obhdp.2008.03.003

Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis - verbal reports as data (revised edition)*. MIT Press.

Jensen, E. (2002). Is explicit information important for performance in dynamic systems? In P. I. Davidsen, E. Mollona, V. G. Diker, R. S. Langer, & J. I. Rowe (Eds.), *Proceedings of the 20th International Conference of the System Dynamics Society* (pp. 1–18). Albany, NY: System Dynamics Society. Retrieved from http://www.systemdynamics.org/conferences/2002/proceed/papers/Jensen1.pdf

Kainz, D., & Ossimitz, G. (2002). Can students learn stock-flow-thinking? An empirical investigation. In P. I. Davidsen, E. Mollona, V. G. Diker, R. S. Langer, & J. I. Rowe (Eds.), *Proceedings of the 20th International Conference of the System Dynamics Society* (pp. 1–34). Albany, NY: The System Dynamics Society. Retrieved from http://www.systemdynamics.org/conferences/2002/proceed/papers/Kainz1.pdf

Kapmeier, F. (2004). Findings From Four Years of Bathtub Dynamics at Higher Education Institutions in Stuttgart. In J. M. Kennedy, Michael; Winch, Graham W.; Langer, Robin S.; Rowe, Jennifer I.; Yanni (Ed.), *Proceedings of the 22nd International Conference of the System Dynamics Society* (pp. 1–22). Albany, NY: System Dynamics Society.

Kasperidus, H. D., Langfelder, H., & Biber, P. (2006). Comparing systems thinking inventory task performance in German classrooms at high school and university level. In A. Groessler, E. A. J. A. Rouwett, R. S. Langer, J. I. Rowe, & J. M. Yanni (Eds.), *Proceedings of the 24th International Conference of the System Dynamics Society* (pp. 1–28). Albany, NY: System Dynamics Society.

Korzilius, H. P. L. M., Raaijmakers, S., Rouwette, E., & Vennix, J. A. M. (2014). Thinking aloud while solving a stock-flow task: Surfacing the correlation heuristic and other reasoning patterns. *Systems Research and Behavioral Science*, *31*(2), 268–279. http://doi.org/10.1002/sres.2196

Lyneis, J. M., & Lyneis, D. A. (2003). Bathtub Dynamics at WPI. In J. D. Sterman (Ed.), *Learning Bathtub Dynamics: A Follow-up*. New York: The 21st International Conference of the System Dynamics Society. Retrieved from http://www.systemdynamics.org/conferences/2003/proceed/PAPERS/S01.pdf

Ossimitz, G. (2002). Stock-Flow-Thinking and Reading stock-flow-related Graphs : An Empirical Investigation in Dynamic Thinking Abilities. *The 20th International Conference of The System Dynamics Society*, (May), 1–26. Retrieved from http://www.systemdynamics.org/conferences/2002/proceed/papers/Ossimit1.pdf

Pala, Ö., & Vennix, J. A. M. (2005). Effect of system dynamics education on systems thinking inventory task performance. *System Dynamics Review*, *21*(2), 147–172. http://doi.org/10.1002/sdr.310

Reed, S. (1999). *Word Problems: research and curriculum reform. Mahwah.* N.J.: Lawrence Erlbaum Associates.

Röder, V., Sedlmeier, P., & Schwarz, M. (under review). Can Gender Priming Eliminate the Effects of Stereotype Threat? The Case of Simple Dynamic Systems.

Schwarz, M., Epperlein, S., Brockhaus, F., & Sedlmeier, P. (2013). Effects of illustrations, specific contexts, and instructions: further attempts to improve stock–flow task performance. In R. Eberlein & I. J. Martínez-Moyano (Eds.), *Proceedings of the 31st International Conference of the System Dynamics Society*. Albany, NY: System Dynamics Society. Retrieved from http://www.systemdynamics.org/conferences/2013/proceed/papers/P1168.pdf

Sedlmeier, P. (1998). The Distribution Matters : Two Types of Sample-Size Tasks. *Journal of Behavioral Decision Making*, *11*, 281–301.

Sterman, J. D. (2002). All models are wrong: Reflections on becoming a systems scientist. *System Dynamics Review*, *18*(4), 501–531. http://doi.org/10.1002/sdr.261

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Cognitive load theory. New York, NY: Springer. doi:10.1007/978-1-4419-8126-4

van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The Think Aloud Method - A practical guide to modelling cognitive processes*. London: Academic Press.

Veldhuis, G., & Korzilius, H. P. L. M. (2012). Seeing with the mind - The role of spatial ability in inferring dynamic behaviour from graphs and stock and flow diagrams. In E. Husemann & D. Lane (Eds.), *Proceedings of the 30th International Conference of the System Dynamics Society* (pp. 1–23). Albany, NY: System Dynamics Society.

Youssef-shalala, A., Ayres, P., & Schubert, C. (2014). Using a General Problem-Solving Strategy to Promote Transfer. *Journal of Experimental Psychology: Applied*, *20*(3), 215–231. http://doi.org/10.1037/xap0000021