Effects of Illustrations, Specific Contexts, and Instructions: Further Attempts to Improve Stock–Flow Task Performance

Marcus A. Schwarz, Saskia Epperlein, Friederike Brockhaus, & Peter Sedlmeier

Chemnitz University of Technology, Germany

Address for correspondence:

Marcus Schwarz

Department of Psychology

Chemnitz University of Technology

09107 Chemnitz

Germany

Email: marcus.schwarz@psychologie.tu-chemnitz.de

Phone: ++49-371-53138948

Fax: ++49-371-531838948

Abstract

Although we face a multitude of complex dynamic systems every day, there is empirical evidence that even simple ones such as stock–flow (SF) systems are extremely difficult to understand. Based on different theoretical approaches and on previous findings in educational and cognitive research, the current study investigated two approaches to improving performance in SF tasks: invoking valid mental models and building new suitable mental models. In two experiments, the effects of net-flow data illustrations, supportive chart representations, selected contextual scenarios, and two adapted educational methods (informative instruction and induced discovery) on SF task performance were empirically tested. Results indicate that none of the approaches led to increased SF task performance. However, gender and mathematical skills were found to be valid predictors of task solutions.

Keywords: *stock–flow, dynamic systems, problem solving, representational formats, instructions*

Facing numerous dynamic systems in quite different situations every day, we are often challenged in handling them and maybe sometimes unable to meet the challenge. In fact, even quite simple dynamic systems, such as a bank account, are complicated enough to cause difficulties for many people. Bank accounts can be seen as a kind of stock–flow (SF) system, because a stock (the account balance) is determined by inflows (revenues) and outflows (expenses) over time. If expenses exceed revenues, sooner or later the account owner will be in trouble.

Even though SF systems lack further features of dynamic systems, such as time delays, feedback loops, and nonlinear relations (Sterman, 2002), there is considerable empirical evidence that SF tasks are very difficult. In numerous studies only about half of the participants solved SF tasks correctly (Booth Sweeney & Sterman, 2000; Brunstein, Gonzalez, & Kanter, 2010; Cronin, Gonzalez, & Sterman, 2009; Gonzalez & Wong, 2011; Ossimitz, 2002; Phuah, 2010; Strohhecker, 2009, 2011). Because most participants were recruited from a pool of well-educated college students, the performance of the average population might be considerably worse. Surprisingly, the rather humble solution rates appeared to be quite resilient, as factors such as graphical literacy, motivation, cognitive capacity, domain familiarity, and contextual embeddedness of the task seemed to have no effect on performance (Brunstein et al., 2010; Cronin & Gonzalez, 2007; Cronin et al., 2009). In sum, the preliminary empirical findings reveal the need for further attempts to improve SF task performance.

To some extent, participants seem to use anchoring or adjustment heuristics to solve SF tasks (Moxnes, 1998; Sterman, 1989). The correlation heuristic (Cronin et al., 2009) the phenomenon that participants tend to assume that changes in the stock correspond to changes in the flow variables. Participants may use the course of inflow, outflow, or net flow (the difference between the flow variables) to generate an answer on the stock’s progression. But of course not every incorrect answer in an SF task can be explained by the correlation heuristic. To improve SF task performance it might be useful to consider the possible processes and strategies, both correct and incorrect, that can be applied in solving SF tasks.

According to the theory of human problem solving (Newell & Simon, 1972), problem solving begins with a mental problem representation. This representation is characterized by the problem statement and the given task information. It enables the problem solver to determine the current state (the given information) and the solution state (the required answer) of a problem. Between these states the so-called problem space contains all possible steps and interim solutions as well as all possible operations to change the state or reach the solution state. To choose a particular operation means to change the state and process the problem. Thus the selection of an operation is primarily determined by the features of the representation. Using certain operations will lead to an altered problem representation, to a termination of processing, to a change of method, or finally to the solution.

Given this basic understanding of human problem solving, the solution of any problem (such as any SF task) relies on the selection of operations. Although this depends on the mental representation of the problem, the result of the problem-solving process is strongly shaped by the problem statement and the given information (e.g., the outflow and inflow information). Thus, it seems plausible that the problem statement, respectively the explanations and the display of information have a considerable impact on the solution rates of SF tasks. Thus, from a problem solving perspective, SF task perception might be improved by using adequate representations, which should trigger existing mental models. Although a suitable problem statement could evoke a suitable mental representation of a given system that might help in solving the respective task, such a mental representation might not exist. In this case, it seems reasonable to encourage and facilitate the development of a (new) mental model. In two experiments, we focused on the two possibilities for improving SF problem solving: invoking valid mental models and building new suitable mental models.

Invoking Valid Mental Models

To design problem statements that activate profound mental representations of SF systems, we investigated the role of providing adequate information, appropriate formats, and suitable context scenarios. We begin with the kind of information that should be provided. The usual way to present SF tasks is to include a description of the system and data on inflow and outflow over time. Most participants seem to understand these data quite well and studies have reported up to 90% correct answers to questions about the progression of the given in- and outflows (Brunstein et al., 2010; Cronin & Gonzalez, 2007; Cronin et al., 2009; Ossimitz, 2002; Sterman, 2002; Strohhecker, 2009). However, the same studies reported quite disappointing rates of correct solutions, indicating a considerable gap between participants' understanding of the given data and their *correct* processing. Of course, considering the inflow and outflow data is only the first step to arrive at the correct solution of an SF task. In further steps, the difference between the flow variables (the so called net flow) has to be (a) identified and (b) accumulated over time. Thus, a first impediment to a correct solution might be that participants do not realize that they have to compile the net flow and, possibly, how to do this.

Presenting the net flow could reduce this first difficulty and thereby increase the solution rate of SF tasks. Providing the net flow should (a) make the problem statement more specific and give more relevant information, (b) reduce the problem space by eliminating potentially wrong interim steps, and (c) reduce the selection of erroneous operations such as the correlation heuristic. This should lead to a more suitable mental representation of the SF task and thereby to a better choice of method, resulting in a more effective solution process.

The way information is provided, and not just the type of information, influences the mental representation of a given problem, as well. If, for example, the layout of controls does not match the actual arrangement of lights in a room, people seem to have considerable difficulty turning on the intended light. Norman (1988) called this effect *affordance*: External representations may prompt spontaneous judgments, which in turn may mislead or support the user's decisions. The way representation contributes to performance on a logical reasoning problem can also be seen in the Wason Selection Task. This is a logic puzzle in which the truth of a proposition is tested by selecting one or more cards from a set of four cards. Wason & Shapiro (1971) found that if the proposition was presented in a rather abstract way (e.g., If there is a vowel on one side of the card, there has to be an even number on the other side), only 12% of the participants selected the correct two cards to check whether the rule was violated. Changing nothing but the context of the task (e.g., If somebody is drinking alcohol, he has to be of legal age) increased the solution rates dramatically (Cosmides, 1989). The concepts of affordance and contextual differences in logical judgments indicate that there might be some kind of preference for the way information is represented that yields improved performance. A similar conclusion emerged from research on the impact of representation formats on statistical reasoning (Sedlmeier, 2007). Performance in Bayesian tasks was considerably improved by providing natural frequencies (e.g., one of four persons) instead of the usual likelihoods (e.g., 25%; Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007; Hoffrage, Gigerenzer, Krauss, & Martignon, 2002; Sedlmeier & Gigerenzer, 2001). Indeed, it seems that how information is represented has a remarkable impact on human decision making in general (Sedlmeier & Hilton, 2012).

Lalomia, Coovert, and Salas (1988) reported that graphical displays facilitated performance in interpolation and trend analysis tasks, whereas numerical presentations were more suitable for value location tasks. Other studies found that students profited from pictures illustrating textual descriptions in terms of learning transfer compared to solely textual descriptions (Mayer, Bove, Bryman, Mars, & Tapangco, 1996; Mayer & Gallini, 1990; Moreno & Mayer, 2000). These benefits of multimedia representations suggest that simultaneous text and illustrations of data should increase performance on SF tasks, as well, especially because such representations support novices even more than experts (Kalyuga, Chandler, & Sweller, 2000; Ollerenshaw, Aidman, & Kidd, 1997).

In sum, there is evidence that not only the kind of information but also the way it is presented affects the processing and solution of tasks. One might assume that adequate information and appropriate representational formats invoke valid mental models of the respective tasks, which leads to a more profound problem statement, which in turn promotes task performance. But these are only two possible ways to invoke valid mental models. Providing suitable context scenarios is another. Earlier studies found no differences in performance across SF task contexts differing in familiarity (Cronin et al., 2009). But familiarity with a given context may differ considerably across participants. The current study considers the suitability of the context scenario, in terms of how well it matches the data. By using an SF task we can distinguish between continuous and discontinuous progressions of the flow variables. According to the matching idea, a different representation will confer an advantage for each of the two progression types: For discontinuous flows, distinct context scenarios (e.g., mentioning humans) might be suitable, whereas for continuous data, liquid context scenarios (e.g., mentioning water) might be advantageous. It seems worthwhile to examine whether matching the contextual representation to the underlying data promotes the processing and solving of SF tasks.

Building Valid Mental Models

The invoking approach focuses on optimizing the kind of information used and the way it is presented. But given the poor SF task performance found in most studies, it is possible that participants lack preexisting mental representations, or at least those they have may be rather limited in their usability. Thus, encouraging and facilitating the development of a valid mental model could be another way to improve SF task performance.

This could be done in two ways: by highlighting relations between variables and drawing attention to the key aspects of SF tasks—which is nothing but prompting a correct solution more or less directly—and by encouraging participants in exploration and discovery—which should promote the autonomous development of a valid mental model. Both approaches can also be framed as *instruction* (prompting a solution) or *discovery* (having participants explore a solution space), frameworks that have been extensively discussed in educational literature and were first described by Bruner (1961). Both approaches have been shown to be relevant for children’s learning in a number of studies (instructions: Csibra & Gergely, 2009; Gergely, Egyed, & Király, 2007; Koenig & Harris, 2005; Kushnir, Wellman, & Gelman, 2008; Tomasello & Barton, 1994; discovery: Schulz

& Bonawitz, 2007; Piaget, 1929; Schulz & Bonawitz, 2007; Singer, Golinkoff, & Hirsh-Pasek, 2008) and both are effective teaching strategies.

SF research could take advantage of both concepts. An *informative instruction* design might highlight the most relevant features of an SF task, whereas an *induced discovery* design might trigger autonomous exploration of the SF system. The usual SF task description includes the problem statement and the data for the flow variables. Yet participants may not instantly recognize what information is important or how to process it adequately. But what are the most relevant features of the data? One way to answer this question is to have a closer look at an *ideal* solution.

Cronin et al. (2009) reported a quite elegant technique for solving an SF task by considering three simple properties of stocks and flows: (A) A stock rises when the inflow exceeds the outflow and it declines when the inflow falls below the outflow. (B) When there is change from one of the two proportions (described in A) to the other there must be a turning point in the stock. If the inflow exceeds the outflow, the stock at the turning point has to be greater than at the beginning (and vice versa if inflow falls below outflow). C) To determine the final stock one needs simply to compare the areas between the rates of inflow and outflow. If the area between the rates is smaller when the inflow exceeds the outflow compared to the area between the rates when the outflow exceeds the inflow, the outflow exceeds the inflow overall, and the final stock has to be smaller than the initial one (due to A). If the area between the rates is smaller when the outflow exceeds the inflow, the final stock has to be greater than the initial stock. Given these simple properties, it is possible to see how a stock develops from the beginning, to a turning point, to the end—without any calculations. Embedding this simple solution strategy in an informative instruction and providing it to participants before they solve ordinary SF tasks might highlight the most relevant features and relations in the data and thereby enhance SF task performance.

A second way to enable participants to build valid mental representations of SF systems might be to encourage them to explore and discover SF systems by themselves, through induced discovery. According to Alfieri, Brooks, Aldrich, and Tenenbaum (2011), an induced discovery approach provides minimal guidance but no explicit target information. In the SF context, individual reflections could be induced, for example, with open questions. Participants should benefit from the autonomous exploration of the SF system, which in turn should increase the system knowledge and thereby enhance SF task performance.

The Current Study

In two experiments we investigated the two approaches for increasing SF task performance: invoking valid mental models and building new suitable mental models. In Experiment 1 we made three attempts to invoke existing mental representations: (1) by presenting adequate net flow information, (2) by providing this information appropriately through text and additional charts, and (3) by using suitable context scenarios (matching data and scenario). Thus, we varied what information was given (with vs. without net flow), how this information was given (text with and without additional charts), and the suitability of the context (matched vs. mismatched to the data). Experiment 2 addressed the effects of instructions and discovery, both adapted from educational approaches, on the facilitation of building new suitable mental models. Throughout both experiments several tasks, differing in degree of difficulty, were used to examine the effects of manipulations on a variant of the tasks and to prevent potential bottom or ceiling effects.

Experiment 1

The purpose of this experiment was to examine how valid mental models might be invoked in order to improve SF task performance. We investigated whether the kind of information that is presented, the way this information is presented, and the suitability of the context scenario contribute to performance. More specifically, we tested three hypotheses: the net-flow hypothesis, the combination hypothesis, and the matching hypothesis. The net-flow hypothesis predicts better SF task performance if the net flow is (a) given or (b) developed, compared to a control condition without net-flow presentation. This should be the case because net flow is an essential piece of information and its presentation/development is expected to make the problem statement more specific, diminish the subjective problem space, and reduce the use of heuristics. The combination hypothesis predicts that the combined presentation of textual and graphic information will increase SF task performance compared to presentation of text only, because multiple information displays should, given the benefits of multimedia representations, foster the encoding and processing of information. The matching hypothesis predicts better SF task performance when the context of the provided scenarios (mall vs. bathtub) matches the characteristics of the underlying data (discontinuous vs. continuous).

To examine the hypotheses we used a 2×4 between-subjects design resulting in eight conditions. We provided two contextual scenarios (mall vs. bathtub) and four kinds of representations (text only vs. text and flow charts vs. text, flow charts and net-flow charts vs. text with flow charts and net-flow development). In each condition we asked each participant to perform three different SF tasks. Whereas the contextual conditions allowed us to test the matching hypothesis, the different representational formats could be used to test both the combination hypothesis and the net-flow hypothesis.

Method

Participants. One hundred and twenty undergraduates from Chemnitz University of Technology (63% female; mean age: 22.5 years, *SD* = 3.1) participated in the experiment and received ether course credit to satisfy an academic requirement or monetary compensation of 4 euros.

Material. The whole experiment was conducted using a paper-and-pencil procedure. The three SF tasks employed in each of the conditions had been used in earlier studies (see Figure 1) and have been found to vary considerably in their degree of difficulty. The W task (W pattern of inflow) was the most difficult; the S task (step pattern of inflow) has been found be less difficult but comparable to the D task (discontinuous in- and outflow). For the two contextual models the tasks were embedded in either a bathtub scenario (water is running in and out of a bathtub) or a mall scenario (persons are entering and leaving a mall).

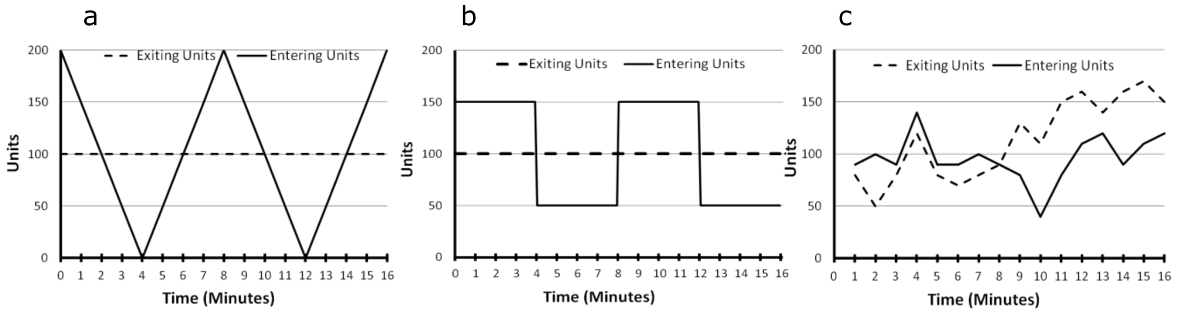

Figure 1. The three stock–flow tasks used in Experiment 1. The W task (a) and the S task (b) were adapted from the bath tub and the cash flow tasks of Booth Sweeney and Sterman (2000). The D task (c) was adapted from the department store task of Sterman (2002).

Procedure. Participants were tested individually. Participants were informed that they were to perform three tasks and were asked to switch off their cell phones and refrain from talking to other participants. Thereafter they were given a booklet that contained the three SF tasks followed by sociodemographic questions. The three SF tasks were presented in random order and, in either the bathtub context or the mall context scenario, illustrated according to one of the illustration conditions (text only vs. text and flow charts vs. text, flow charts, and net-flow charts vs. text with flow charts and net-flow development). Given the inflow and outflow information of these tasks (see Figure 1), participants were then asked to draw the estimated course of the stock over the respective period of time on an empty chart. Additionally, they had to answer two questions about the times of the greatest inflow and greatest outflow. Additional questions asked about the time of the greatest stock and the smallest stock. Finally we collected demographic data, such as gender, age, final grade in mathematics at high school and final high school exam.

Measurement. The participants' drawings of the estimated course of the stock were rated by two independent raters, following criteria introduced by Booth Sweeney and Sterman (2000). With slight modification we arrived at five criteria for the rating of participants' drawings: stock rises during positive net flow, stock decreases during negative net flow, stock reaches maximum or minimum when the net flow is zero, no jumps in the course of the stock, and the rate of the slope of the stock reflects the course of the net flow. Thus, each stock drawing yielded 0 to 5 points from each rater. Points from both raters were summed and converted into percentages. The four further questions on the flow and stock extremes were scored with either 0 points for wrong answers or 1 point for correct answers. Effect sizes are reported by Hedges's *g* (Rosenthal, Rosnow, & Rubin, 2000; Sedlmeier & Renkewitz, 2008) or by eta squared.

Results

In sum, the SF tasks turned out to be still difficult to solve, despite adapted representation formats. Overall, the participants received on average 51.7% (*SD* 31.1%) of the maximally possible points. The solution rates for the three tasks varied considerably ($M_{\text{W task}}$ = 28.9%, *SD* = 39.7%; $M_{\text{S task}}$ = 61.3%, *SD* = 43.3%; $M_{\text{D task}}$ = 64.8%, *SD* = 34.4%). In answering the two questions on the stock extremes the participants reached overall similar results by 50.6% of the maximally possible points (*SD* 30.9%; $M_{\text{W task}}$ = 26.3%; *SD* = 42.0%; $M_{\text{S task}}$ = 46.7%; *SD* = 42.4%; $M_{\text{D task}}$ = 78.8%; *SD* = 37.0%). Both dependent measures correlated highly positively, r_{overall} = .88 ($r_{\text{W task}}$ = .83; $r_{\text{S task}}$ = .76; $r_{\text{D task}}$ = .80).

The presentation and development of the net flow were expected to make the problem statement more specific, diminish the subjective problem space, and reduce the use of heuristics. Thus, the net-flow hypothesis holds that SF task performance should increase when the net flow is given or developed compared to no net-flow presentation. Figure 2 illustrates the mean SF task performance by presentation condition.

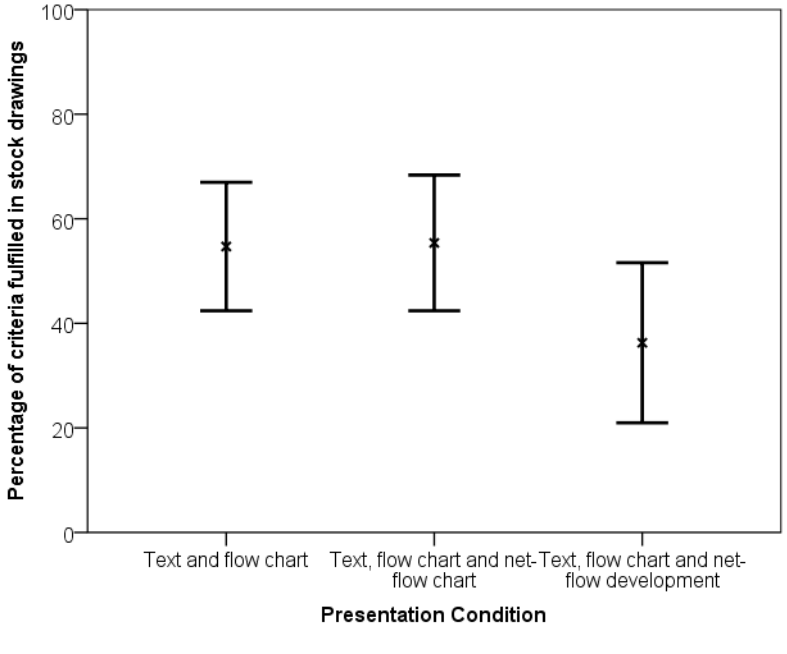

Figure 2: Mean stock–flow task performance as percentage of criteria fulfilled over the three experimental conditions: text and flow chart; text, flow chart, and net-flow chart; and text, flow chart, and net-flow development. Error bars represent 95% confidence intervals.

As we found the expected performance differences between the three tasks, no relevant differences between the conditions occurred for any of the tasks. Therefore, in the following we report only total performance aggregated over all tasks (for task-specific results please see Appendix A). The mean SF task performance for net-flow development was the lowest among the three conditions in Figure 2 ($M_{\text{text + flow chart}}$= 54.7%; $M_{\text{text, flow chart, + net-flow chart}}$ = 55.4%; $M_{\text{text, flow chart, + net-flow development}}$= 36.3%). None of the differences reached significance, $F(2,69)=2.72$, $p= .073$, η²= .07. Because net-flow development may not lead automatically to correct net-flow pattern, we coded the sketched charts in this condition on their correctness. Sixty-seven percent of the participants compiled all net flows correctly. Although these participants were slightly better performing ($M_{\text{correct net-flow development}}$ = 40.1%), we still found no relevant differences nor any support for the net-flow hypothesis, $F(2,69) = 1.23$, $p = .299$, η²= .03.

Because multiple information displays should foster the encoding and processing of information, the combination hypothesis predicts an increase in SF task performance for the combined presentation of textual and graphic information compared to solely textual representations. Figure 3 shows the mean SF task performances in the condition with only textual presentation and those with additional chart presentations.

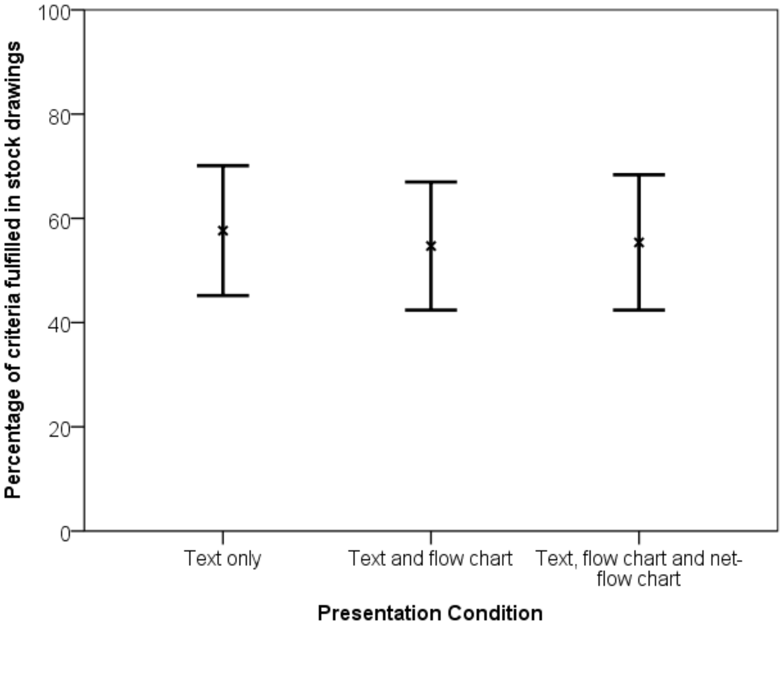

Figure 3: Mean stock–flow task performance as percentage of criteria fulfilled over three experimental conditions: text only, text and flow chart, and text with flow chart and net-flow chart. Error bars represent 95% confidence intervals.

Inconsistent with the combination hypothesis, the mean total SF task performance[1] for solely textual presentation was higher than that in the other presentation conditions ($M_{\text{text only}}$ = 57.6%; $M_{\text{text + flow chart}}$ = 54.7%; $M_{\text{text, flow chart, + net-flow chart}}$ = 55.4%). However, none of the differences reached significance (F(2,69)=0.64, p= .938, η²= .02).

We expected better SF task performance if the context scenario was matched to the characteristics of the underlying data (continuous vs. discontinuous data). According to the matching hypothesis*,* the discontinuous D task was expected to be performed better within a scenario that was based on distinct entities (people in a mall) compared to a continuous scenario (water in a bathtub). However, the performance in the W or S task should benefit from a continuous scenario and decrease within a discontinuous one. Figure 4 shows the mean SF task performances for the W, S, and D tasks dependent on the contextual conditions mall versus bathtub scenario (people vs. water).

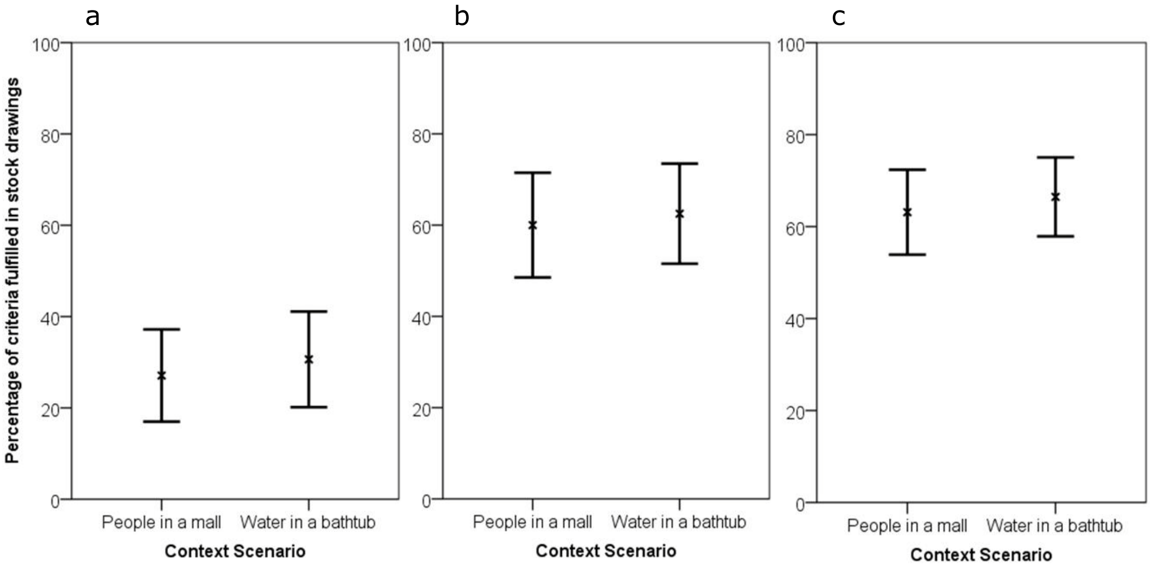

Figure 4: Mean stock–flow task performance as percentage of criteria fulfilled by context condition (people in a mall or water in a bathtub), for (a) the W task, (b)

[1] For task-specific results please see Appendix A.

the S task, and (c) the D task (all representational formats aggregated). Error bars represent 95% confidence intervals.

Performance in the rather continuous S and W tasks indeed increased in the bathtub context (match of context and data), but performance in the discontinuous D task increased as well in the same context (mismatch of context and data). However, none of the differences between the contexts within the tasks were found to be substantial, $t_{\text{W task}}(118) = 0.49$, $p = .627$, $g = 0.09$; $t_{\text{S task}}(118) = 0.32$, $p = .753$, $g = 0.06$; $t_{\text{D task}}(118) = 0.53$, $p = .598$, $g = 0.10$.

For further explorative analyses the demographic data were subjected to a regression analysis regarding impact on mean SF task performance. We found effects of gender (advantages for males) and grades in mathematics but none of age and general high-school exam on mean SF task performance ($ß_{\text{gender}}$=.315, p=.001; $ß_{\text{mathematics}}$=.214, p=.043; $ß_{\text{age}}$=.044, p=.632; $ß_{\text{high-school exam}}$=-.106, p=.373).[2] Examining possible interactions of gender and mathematical skills we obtained partial correlations showing moderate effects ($r_{\text{gender × SF task perform., part mathematics}}$= .33, p<.001; $r_{\text{mathematics × SF task perform., part gender}}$ = .31, p=.001).

Discussion

Experiment 1 explored three attempts to invoke valid mental models and thereby increase SF task performance. The net-flow hypothesis and the combination hypothesis predicted improvement in performance if adequate information was provided in an appropriate way. The matching hypothesis predicted improvement when the context scenario matched the characteristics of the underlying data (discontinuous vs. continuous). In sum, the results of Experiment 1 did not support the hypotheses. None of the experimental variations had a distinct impact on SF task performance.

The net-flow hypothesis assumes that providing or developing the net flow, as essential information, will raise SF task performance. We found no substantial differences

[2] Note that the reported ß values are standardized regression weights reflecting the particular effect of each factor on SF task performance while simultaneously controlling for all other factors. The negative ß for high-school exams is a result of the way grades are coded in German schools (1–5, with 1 being very good and 5 insufficient). The grade in mathematics is expressed on a point scale (0–15, insufficient to very good).

between the conditions of text and flow chart; text, flow chart, and net-flow chart; and text, flow chart, and net-flow development. Surprisingly, the SF task performance was even lower when the net flow had to be compiled first, compared to the text only condition, *t*(46)=2.24 *p*=.030, *g*=0.65. One possible reason might be information overload, which could have impaired participants' cognition in general. Another reason might be that the presentation of the flow charts led to misperceptions. The flow charts illustrate rates (e.g., liters per minute), but the stock to be drawn in the stock chart required an absolute amount (e.g., a liter). Because the two charts look very similar but rely on different units, participants may have misinterpreted one or more of the charts. The combination hypothesis predicts an increase of SF task performance for combined textual and chart presentations, but we found no differences between these conditions and the text only condition. However, we found no advantage for the solely textual representation, either, although such an advantage has been reported elsewhere (Fischer & Degen, 2012). Again, one reason for these results might be information overload. This may have interfered with possible positive effects of the combined presentation, because performance was very similar in the compared conditions. The matching hypothesis predicts better SF task performance if the context scenario matches the data characteristics. We also found no support for this hypothesis.

Although—or even because—none of the conditions showed improvement in SF task performance, the overall results for the rather difficult W task (29%) and the relatively simple S task (61%) illustrate the high degree of difficulty of SF tasks. In this respect, the current results are consistent with numerous findings of previous studies, as mentioned in the Introduction.

The investigation of the sociodemographic variables revealed moderate effects of gender and mathematical skills on SF task performance, with males and mathematically skilled participants seeming to have an advantage. Although there are rather inconsistent findings on the impact of gender (e.g., Booth Sweeney & Sterman, 2000; Kainz & Ossimitz, 2002; Ossimitz, 2002; Sterman, 2010), the degree of mathematical education has often been thought to affect SF task performance (e.g., Booth Sweeney & Sterman, 2000; Cronin et al., 2009; Kainz & Ossimitz, 2002; Kapmeier, 2004). In previous studies, mathematical education was often assessed rather qualitatively, such as by means of postsecondary field of study (e.g., Booth Sweeney & Sterman, 2000; Kapmeier, 2004). These categorial measurements might be vulnerable to biases. For example, not everyone who chooses mathematics as a major needs to be outstanding in mathematics. Likewise, being educated or employed in a nonmathematical field does not necessarily mean one is weak in mathematics. Therefore in the current experiment we measured mathematical skills as the final high-school grade in mathematics. This enabled us to examine the relation of mathematical skills and SF task performance more precisely.

Because none of our attempts to invoke valid mental models of SF systems improved SF task performance, building new suitable mental models might be more effective than (only) providing adequate representation formats to invoke possibly wrong (or nonexistent) mental models. Two possible ways to build or specify new mental models of SF systems are *instruction* and *discovery*. Both concepts have been shown to be effective teaching strategies in educational research. We adapted them to SF problems and investigated their impact in Experiment 2.

Experiment 2

The previous experiment as well as earlier studies tried to optimize presentational and contextual formats. In contrast, in Experiment 2 we pursued the idea of building or refining a suitable mental representation to solve SF problems. We investigated the impact of four conditions on the performance in SF tasks: an informative instruction condition, an induced discovery condition, and two baseline conditions, a semi-informative instruction condition and a control condition. The informative instruction condition was derived from an idealized solution of Cronin et al. (2009), who reported an elegant technique to solve an SF task with few simple considerations (see above). In the induced discovery condition, the relevance of the relation of inflow and outflow is indicated by an open question, which should induce exploration and thereby promote self-generated elaboration of the task. The semi-informative instruction condition consisted of questions about the given inflow and outflow and is thus comparable to baseline conditions reported elsewhere (Cronin & Gonzalez, 2007; Cronin et al., 2009). As such it can be considered a second control condition, in addition to the control condition in which no questions were asked or hints provided. The four conditions were implemented in a between-subjects design. We expected that the instructions would enhance SF task performance compared to the control conditions. Thus, the informative instruction hypothesis holds that prompting the crucial steps of the solution by questioning the participants should increase SF task performance compared to the two baseline conditions. The induced discovery hypothesis holds that encouraging participants to explore the SF system themselves by asking an open question should enhance SF task performance compared to both baseline conditions.

Method

Participants. Eighty undergraduates of the Chemnitz University of Technology (45% female; mean age: 22.0 years, *SD* = 2.8) participated in the experiment and received either course credit to satisfy an academic requirement or monetary compensation of 4 euros.

Material. The experiment was conducted as a paper-and-pencil procedure. The two SF tasks used were similar to the S task and the D task of Experiment 1 and were also derived from tasks used in previous studies (see Figure 5). The two tasks showed similar degrees of difficulty in Experiment 1. However, in Experiment 2 both tasks were altered regarding the length of the time span (12 instead of 16 min), and the S task included only the first half of the original S task pattern. This modification was aimed at making the two tasks comparable as the flows cross once in both tasks (at Minute 6 in the S task and Minute 7 in the D task). The context scenario for both tasks was people entering and leaving a cafeteria.

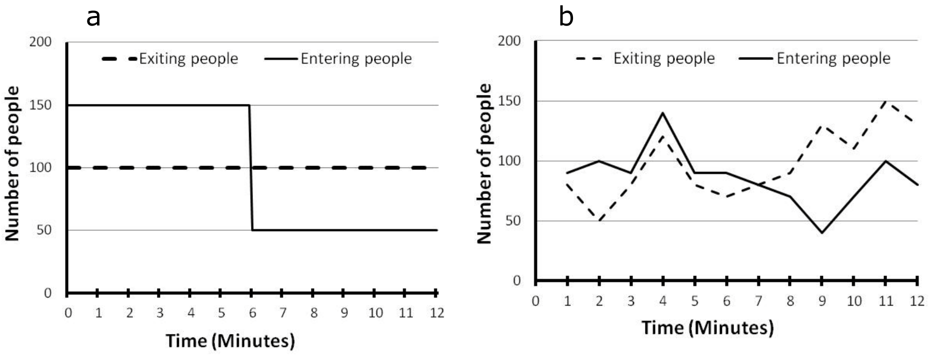

Figure 5: (a) The S task and (b) the D-task used in Experiment 2.

Procedure. Participants were again tested individually, asked to switch off their cell phones, asked to refrain talking to other participants, and informed that they were to perform two tasks. The booklet that was given to them contained the two SF tasks followed by sociodemographic questions about their age, gender, final grade in mathematics at high school and final high school exam. Prior to both SF tasks the participants obtained the instructions pertaining to one of the four conditions. In the informative instruction condition, participants had to answer eight questions indicating the simple stepwise solution described above (see Appendix B). In the induced discovery condition, participants were encouraged to explore the SF task by answering an open question (Which components of the flow diagram seem to be crucial to you and how are they related to each other?). In the semi-informative instruction condition participants had to answer two questions about the point in time at which inflow and outflow were at their maxima. In the control condition the participants got no instructions, nor any additional information. To solve the SF tasks, participants were again asked to draw the changes in the stock over the respective period of time on an empty chart. At the end of the procedure participants had to provide the same demographic data as in Experiment 1.

Data analysis. The participants' drawings of the estimated course of the stock were rated by two independent raters according the same criteria described in Experiment 1. Scores were likewise summed and converted into percentages.

Results

Overall, the participants received on average 48.4% (*SD* = 37.3%) of the maximally available points. The solution rates for the two tasks differed markedly ($M_{\text{S task}}$= 61.4%; $M_{\text{D task}}$= 35.5%), *t*(79)=2.23, *p*<.001, *g*=0.70.

Because the two tasks differed unexpectedly in participants' performance, we examined the impact of the conditions separately for the two tasks. Figure 6 shows the mean SF task performance of the four experimental conditions.

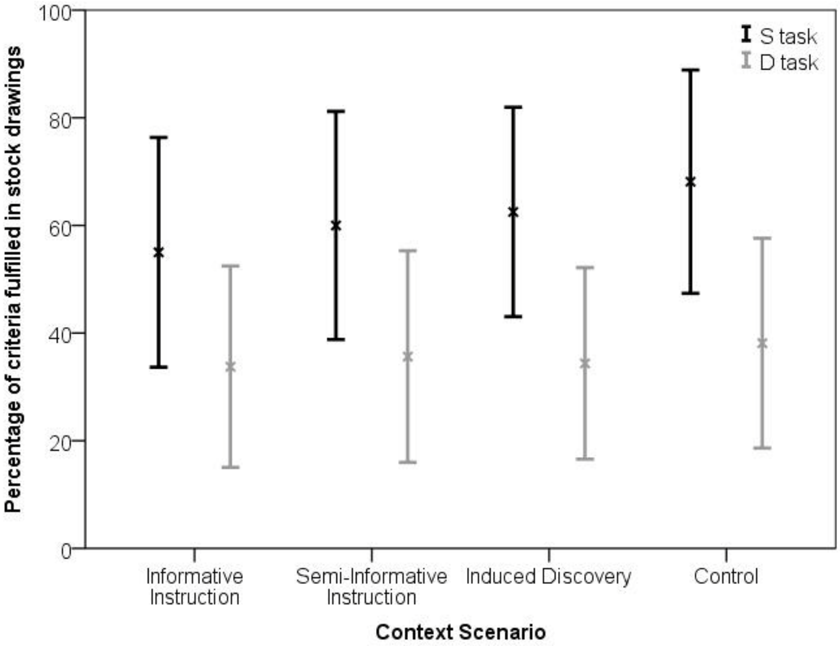

Figure 6: Mean stock–flow task performance as percentage of criteria fulfillment for S tasks and D tasks in Experiment 2, by presentation condition. Error bars represent 95% confidence intervals.

The SF task performance did not vary substantially between the conditions—neither for S tasks nor for D tasks, $F_{\text{S tasks}}(3,76)=0.31$, $p= .822$, $\eta^2= .01$; $F_{\text{D tasks}}(3,76)=0.05$, $p= .987$, $\eta^2< .01$. Also, we found no difference between the control condition and the baseline condition of previous studies (our semi-informative instruction condition), $t_{\text{S task}}(38)=-0.57$, $p=.570$, $g=-0.18$; $t_{\text{D task}}(38)=-0.19$, $p=.851$, $g=-0.06$.

Comparing the informative instruction with the control condition we found no support for the informative instruction hypothesis, $t_{\text{S task}}(38)=-0.92$, $p=.362$, $g=-0.29$; $t_{\text{D task}}(38)=-0.34$, $p=.737$, $g=-0.11$. Note that the negative effect size refers to an actual decrease in SF task performance in the informative instruction condition. Because one might argue that the informative instruction did not affect every participant equally, we analyzed the participants' responses to the informative instruction questions. Comparing only those participants who were at the median level or above of correct responses and the control condition participants, we still found no support for the informative instruction hypothesis, $t_{\text{S task}}(31)=-0.17$, $p=.866$, $g=-0.06$; $t_{\text{D task}}(29)= 0.73$, $p=.470$, $g=0.27$.

Comparing the induced discovery with the control condition we found no support for the induced discovery hypothesis, $t_{\text{S task}}(38)=-0.41$, $p=.681$, $g=-0.13$; $t_{\text{D task}}(38)=-$

0.30, *p*=.768, *g*=-0.09. Note that again, the negative effect size refers to an actual decrease in SF task performance in the induced discovery condition.

The impact of demographic variables on SF task performance was again analyzed with the help of a regression analysis. For averaged SF task performance (S task and D Task) we found effects of gender (advantage for males) and grade in mathematics but none of age and general high-school exam on mean SF task performance ($ß_{\text{gender}}$=.610, *p*<.001; $ß_{\text{mathematics}}$=.264, *p*=.009; $ß_{\text{age}}$=-.051, *p*=.546; $ß_{\text{high-school exam}}$=-.138, *p*=.154). Partial correlations to SF task performance revealed strong and independent effects of gender and mathematical skills ($r_{\text{gender × SF task perform., part mathematics}}$= .62, *p*<.001; $r_{\text{mathematics × SF task perform., part gender}}$ = .45, *p*<.001).

Discussion

In Experiment 2 we examined two methods for building new suitable mental models: an instruction and a discovery approach. Although both strategies have found broad support in educational research, the results of Experiment 2 suggest that neither had a relevant impact on our SF task performance. Neither the informative instruction hypothesis nor the induced discovery hypothesis was supported by our data. Even those participants who responded correctly to the instructive questions in the informative instruction condition (at the median level or above) obviously did not benefit in the expected way. Although the necessary puzzle pieces were metaphorically on the table, the participants may not have been able to assemble them correctly. Thus, a more direct or conditional instruction in an if–then format might have been more successful than the indirect way we chose here. On the other hand, the exploration we tried to induce with the open questions in the induced discovery condition might not have been strong enough to evoke sustainable models. It is conceivable that the questions were not helpful or the participants' motivation might have been too low to trigger an exhaustive or at least deeper exploration of the scenario. However, we cannot rule out the possibility that both kinds of instructions failed in leading participants to build adequate representations of the SF tasks.

Of course one has to consider the temporal constraints of the experimental procedure and it is possible that new suitable mental models cannot be established in less than half an hour. It is further conceivable that beneficial and detrimental effects of the instructions cancelled each other out, or that the conditions affected participants differently (positively and negatively) but on a mean level not substantially. The moderate performance on the S task (61%) is consistent with prior empirical findings and with the results from Experiment 1. The low performance on the D task (36%) was, however, surprisingly worse than earlier findings. In this modified version of the department store task it might have been more difficult to determine which overall net flow was dominant (the positive or the negative), because the difference was much smaller than in the original 16-min version. Further, we again found independent effects for gender and mathematical skills, which indicates that the processing of SF tasks depends at least partly on certain personal attributes.

General Discussion

The focus of the present studies was to enhance SF task performance by two main approaches: invoking valid mental models and building new suitable mental models. In Experiment 1 we investigated the first approach by providing or having participants develop net-flow data, by representing this data with text alone or text plus additional charts, and by matching the context scenario to the data characteristics. In Experiment 2 we further investigated the effects of building new suitable mental models by either instructive solution cues or induced discovery. SF task performance in both experiments was measured by independent judges' ratings of participants' drawings of the estimated course of the stock.

In sum, none of the different experimental conditions showed the expected improvement. Neither providing and compiling the net flow, nor combining representations, nor matching contextual representation and data showed the expected benefits. This is probably the most surprising result of Experiment 1, because realizing the net flow was assumed to be an essential step in SF comprehension. Two promising educational methods, instruction and discovery, were obviously ineffective in increasing

SF task performance—as the results of Experiment 2 show. Although the lack of improvement was disappointing, the results corroborate previous empirical findings in the field of SF research: SF tasks are quite difficult and rather resilient to various optimizing strategies.

Many of the approaches to improving participants' performance in SF tasks, in earlier studies as well as in the present one, can be interpreted as attempts to provide better information illustration. This perspective implies that people do have adequate models of the respective SF systems that just need to be "fed" in an appropriate way. Yet, because almost all of the attempts failed to enhance SF performance, it is plausible that the mental models people have of SF systems are far less specific, sufficient, or correct than initially assumed. Following this idea, we tried to facilitate system knowledge and build on participants' mental models about SF systems by inducing individual explorative discovery. Although the results of our second experiment did not meet these expectations, we still believe that supporting the creation of valid mental models is a promising approach to help people solve SF problems, for two reasons.

First, the adaptation of the two educational approaches (instruction and discovery) might have been suboptimal. It needs to be refined and checked for successful manipulations in future experiments. Second, a few findings indicate slight improvements in SF task performance if an illustration of the respective context system is presented (Brockhaus & Sedlmeier, 2013). It is conceivable that such an illustration promotes the mental model of the underlying contextual system by activating implicit knowledge about the attributes and relations of this system, such as when a picture of a bathtub automatically triggers knowledge about how the behavior of water in a tub depends on the inflow and outflow. Although the illustrations are usually static for mostly dynamical systems (e.g., a bathtub), it might be reasonable to investigate the potential of dynamic representations or animations. Animated system simulations could enable people not only to use existing models but also to generate new valid and dynamic ones, which would be a promising approach for future research.

Despite the lack of performance improvements in the attempts described above, we found moderate to strong effect sizes for gender and mathematical skills affecting SF task performance in both experiments. Gender differences have been reported elsewhere (e.g., Booth Sweeney & Sterman, 2000; Kainz & Ossimitz, 2002; Ossimitz, 2002; Sterman, 2010) and possible reasons are being sought in various fields. For instance, neurophysiological findings indicate different brain areas are involved in problem solving for males and females (Speck et al., 2000; Weiss et al., 2003). Explicit or implicit gender stereotypes are thought to facilitate or be detrimental to problem-solving capabilities (McGlone & Aronson, 2006; Sterman, 2010). Moreover, there might be additional factors that could be confounded by gender, such as computer experience (Wittmann & Hattrup, 2004) or interest in problem contexts (Su, Rounds, & Armstrong, 2009), which might lead to gender differences in problem solving or SF task performance. Mathematical education was suspected to contribute to SF task performance by many researchers, but was usually measured by qualitatively, for example by postsecondary field of study (e.g., Booth Sweeney & Sterman, 2000; Kapmeier, 2004). In addressing mathematical skills by a quantitative measurement (final grade in mathematics at high school) we showed that there seems to be a sustainable and independent (of gender) impact on SF task performance. Of course this procedure could be further refined. Because of examination stress or temporal distance from the high-school exam, a more recent measurement of mathematical skills would be preferable. Whatever mechanisms are assumed to underlie the effects of gender and mathematical skills, it seems to be clearly there are differences in the processing of SF systems. Thus, in order to provide an optimized description of SF systems, it seems reasonable to consider different kinds of representational displays to address different groups of participants.

It seems more than likely that economic, environmental, and social systems become more complex and dynamic. Although SF systems are present in many aspects of our daily lives, the current study shows that they are obviously not easy to understand. As the approaches described in the two experiments remain unsuccessful in improving the comprehension of SF systems, they get in line with numerous empirical studies illustrating the big challenge of getting those systems understood. Further efforts to investigate how people understand such systems and to help them do so more successfully are needed. We think facilitating the building of new mental SF representations with animated illustrations and taking obviously solid factors, such as gender and mathematical skills, into account are promising directions for future research.

References

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology, 103*, 1-18.

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

Brockhaus, F., & Sedlmeier, P. (2013). *Is intuition the key to understanding stock and flow systems? The influence of modifying the systems' representation format on the stock flow performance.* Unpublished manuscript, Chemnitz University of Technology. Chemnitz.

Bruner, J. S. (1961). Act of discovery. *Harvard Educational Review, 31*, 21-32.

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

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31*, 187-276.

Cronin, M. A., & Gonzalez, C. (2007). Understanding the building blocks of dynamic systems. *System Dynamics Review, 23*, 1-17.

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

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences, 13*, 148-153.

Fischer, H., & Degen, C. (2012, July). *Stock-flow failure can be explained by the task format*. Paper presented at the 30th International Conference of the System Dynamics Society, St. Gallen, Switzerland.

Gergely, G., Egyed, K., & Király, I. (2007). On pedagogy. *Developmental Science, 10*, 139-146.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest, 8*, 53-96.

Gonzalez, C., & Wong, H. (2011). Understanding stocks and flows through analogy. *System Dynamics Review*, *28*, 3-27.

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition, 84*, 343-352.

Kainz, D., & Ossimitz, G. (2002, July/August). *Can students learn stock-flow-thinking? An empirical investigation.* Paper presented at the 20th International Conference of the System Dynamics Society, Palermo, Italy.

Kalyuga, S., Chandler, P., & Sweller, J. (2000). Incorporating learner experience into the design of multimedia instruction. *Journal of Educational Psychology, 92*, 126.

Kapmeier, F. (2004, July). *Findings from four years of bathtub dynamics at higher management education institutions in Stuttgart.* Paper presented at the 22nd International Conference of the System Dynamics Society, Oxford, England.

Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development, 76*, 1261-1277.

Kushnir, T., Wellman, H. M., & Gelman, S. A. (2008). The role of preschoolers' social understanding in evaluating the informativeness of causal interventions. *Cognition, 107*, 1084-1092.

Lalomia, M. J., Coovert, M. D., & Salas, E. (1988). Problem solving performance and display preference for information displays depicting numerical functions. *ACM SIGCHI Bulletin, 20,* 47-51.

Mayer, R. E., Bove, W., Bryman, A., Mars, R., & Tapangco, L. (1996). When less is more: Meaningful learning from visual summaries of science textbook lessons. *Journal of Educational Psychology, 88*, 64.

Mayer, R. E., & Gallini, J. K. (1990). When is an illustration worth ten thousand words? *Journal of Educational Psychology, 82*, 715.

McGlone, M. S., & Aronson, J. (2006). Stereotype threat, identity salience, and spatial reasoning. *Journal of Applied Developmental Psychology, 27*, 486-493.

Moreno, R., & Mayer, R. E. (2000). Engaging students in active learning: The case for personalized multimedia messages. *Journal of Educational Psychology, 92*, 724.

Moxnes, E. (1998). Overexploitation of renewable resources: The role of misperceptions. *Journal of Economic Behavior and Organization, 37*, 107-127.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Oxford, England: Prentice-Hall.

Norman, D. A. (1988). *The psychology of everyday things*. New York, NY: Basic Books.

Ollerenshaw, A., Aidman, E. V., & Kidd, G. (1997). Is an illustration always worth ten thousand words? Effects of prior knowledge, learning style and multimedia illustrations on text comprehension. *International Journal of Instructional Media, 24*, 227-238.

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

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

Phuah, T. (2010, July). *Can people learn behaviours of stock and flow using their ability to calculate running total? An experimental study.* Paper presented at the 28th International Conference of the System Dynamics Society, Seoul, Korea.

Piaget, J. (1929). *The child's conception of the world*. Oxford, England: Harcourt, Brace.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. New York, NY: Cambridge University Press.

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology, 43*, 1045-1050.

Sedlmeier, P. (2007). Statistical reasoning: Valid intuitions put to use. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 389-419). Mahwah, NJ: Erlbaum.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General, 130*, 380-400.

Sedlmeier, P., & Hilton, D. J. (2012). Improving judgment and decision making through communication and representation (PSYNDEXalert). In M. K. Dhami, A. Schlottmann, & M. R. Waldman (Eds.), *Judgment and decision making as a skill: Learning, development and evolution* (pp. 229-257). New York, NY: Cambridge University Press.

Sedlmeier, P., & Renkewitz, F. (2007). *Research methods and statistics for psychology*. Munich, Germany: Pearson Education.

Singer, D. G., Golinkoff, R. M., & Hirsh-Pasek, K. (2006). *Play = learning: How play motivates and enhances children's cognitive and social-emotional growth*. New York, NY: Oxford University Press.

Speck, O., Ernst, T., Braun, J., Koch, C., Miller, E., & Chang, L. (2000). Gender differences in the functional organization of the brain for working memory. *Neuroreport, 11*, 2581-2585.

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

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

Sterman, J. D. (2010). Does formal system dynamics training improve people's understanding of accumulation? *System Dynamics Review*, 26, 316-334.

Strohhecker, J. (2009, July). *Does a better understanding of accumulation indeed predict a higher performance in stock and flow management?* Paper presented at the 27th International Conference of the System Dynamics Society, Albuquerque, NM.

Strohhecker, J. (2011, July). Illuminating the *Logic of stock management failure—How much does the (mis)understanding of accumulation explain?* Paper presented at the 29th International Conference of the System Dynamics Society, Washington, DC. Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin, 135*, 859-884.

Tomasello, M., & Barton, M. (1994). Learning words in nonostensive contexts. *Developmental Psychology, 30*, 639.

Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology, 23*, 63-71.

Weiss, E., Siedentopf, C. M., Hofer, A., Deisenhammer, E. A., Hoptman, M. J., Kremser, C. (2003). Sex differences in brain activation pattern during a visuospatial cognitive task: A functional magnetic resonance imaging study in healthy volunteers. *Neuroscience Letters, 344*, 169-172.

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

Appendix A

*Experiment 1: Means (*M*) and Standard Deviations (*SD*) for Each Kind of Task for All Conditions.*

Condition	Task	Context scenario			
		People in a mall		Water in a bathtub	
		M	*SD*	*M*	*SD*
Text only	W	17.71	25.26	51.04	46.61
	S	65.63	41.67	78.13	37.74
	D	70.83	29.36	62.50	28.70
Text and flow chart	W	36.46	45.99	23.96	37.10
	S	77.08	38.00	57.29	46.30
	D	79.17	15.39	54.17	43.08
Text, flow chart, and net-flow chart	W	16.67	34.27	43.75	44.11
	S	56.25	50.14	70.83	43.74
	D	66.67	33.00	78.13	25.63
Text, flow chart, and net-flow development	W	38.54	48.11	15.63	36.59
	S	37.50	47.07	44.79	44.42
	D	28.13	40.29	52.08	38.74

Appendix B

The eight questions, indicating the simple stepwise solution, participants had to answer in the informative instruction condition of Experiment 2.

1. When are the most people entering the cafeteria?
O in (the) minute(s) ……
O from minute…... to minute……
O cannot be exactly determined, because ……

2. When are the fewest people entering the cafeteria?
O in (the) minute(s) ……
O from minute…... to minute……
O cannot be exactly determined, because ……

3. Are consistently more people entering than leaving the cafeteria?
O yes O no O do not know

4. Are consistently more people leaving than entering the cafeteria?
O yes O no O do not know

5. Is there a moment at which the numbers of entering and leaving people are identical?
O yes O no O do not know

6. Which total number of people is taller, those entering over all times or those leaving the cafeteria?
O the number of people entering
O the number of people leaving
O the two numbers are equal
O do not know

7. When are the most people in the cafeteria?
O in (the) minute(s) ……
O from minute…... to minute……
O cannot be exactly determined, because ……

8. When are the fewest people in the cafeteria?
O in (the) minute(s) ……
O from minute…... to minute……
O cannot be exactly determined, because ……