Mathematical Self-Efficacy as a Determinant of Successful Learning of Mental Models From Computerized Materials

Cyril Brom and Filip Děchtěrenko
Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic
brom@ksvi.mff.cuni.cz
filip.dechterenko@gmail.com

Abstract: Computerized animations, simulations and games are useful tools for supporting acquisition of mental models. Various personal characteristics, such as prior knowledge and spatial abilities, can influence, in various ways, effectivity of learning from these materials. In comparative studies with between-subject design that investigate learning effects of these materials, it is important to control for these variables because they should be taken as covariates in case the two (or more) research groups are not sampled equally (which happens even in the case of random assignment of participants to the research groups). In addition, it would be useful to have interventions that measure these variables with as few items as possible; to avoid unbearably long questionnaires. In this initial exploratory study we investigate if mathematical self-efficacy, measured by a single question, and self-assessed ability of acquiring mental models (SAAMM), also measured by a single question, predicts learning outcomes; as concerns mental models acquisition. Re-analyzing data from our four recent studies on one of the well-known principles of multimedia learning, personalization principle (N = 75, 85, 76, 41; college students with diverse background), we show that mathematical self-efficacy and SAAMM are moderately correlated (r = .32 - .40) and indeed related to learning outcomes, measured by transfer tests (r = .22 - .57 and .28 - .48, respectively). However, the reasons behind these relationships seem to be complex and diverse, and at least partly dependent on treatments' characteristics. For a complex simulation using graphs and resembling an educational computer game, this relationship can be, to a large extent, explained by mutual relationships between graphing skills, frequency of game-playing, mathematical self-efficacy, SAAMM, and learning outcomes. For a short animation on an electrophysical topic, it can be explained by mutual links between prior electrophysical knowledge, mathematical self-efficacy, SAAMM, and learning outcomes. Only for a short animation on a math/physics-unrelated topic, we could not explain the relationship between mathematical self-efficacy, SAAMM, and learning outcomes by a third variable (however, the graphing test was not administered in this case). In general, this study indicates that our two questions for assessing mathematical self-efficacy and SAAMM are promising instruments for measuring variables that should be controlled for in studies on learning effects of computerized materials with between-subject design, but more research is needed to pin down details.

Keywords: mental models, mathematical self-efficacy, learning outcomes, animations, simulations, serious games

1. Introduction

Advantages of well-developed computerized animations and simulations for supporting construction of new mental models are well known (e.g., Mayer, 2009; Papert, 1980; Wouters et al., 2013). The process of learning complex knowledge representations (rather than facts or skills) can be conceptualized within the Cognitive Theory of Multimedia Learning (Mayer, 2009): construction of new knowledge happens through processing of incoming information, its organizing into meaningful temporal representations in working memory and integration of these representations with prior knowledge structures in long-term memory. This process is influenced in various ways by learners' characteristics, such as spatial abilities (Höffler, 2010) or prior knowledge (Mayer, 2009).

Both mathematical abilities and their self-assessments predict performance in certain complex skill acquisition tasks (e.g., Ackerman et al., 1995; see also Hacket & Betz, 1989), but studies demonstrating analogical results concerning mental models acquisition are unknown to us. Yet one can speculate that persons who readily manipulate abstract mathematical objects in their working memory can be better at manipulating any (non-verbal) mental object in their working memory, which is, according to the Cognitive Theory of Multimedia Learning, a learning advantage. Mathematical self-efficacy (MATHSE) can thus be one of determinants of mental models acquisition. Besides theoretical implications, if we measure MATHSE with a single question and find a MATHSE—learning outcomes relationship, we would have in hands a simple instrument for predicting test scores. This would bring researchers a substantial aid in comparative studies with between-subject design that investigate learning effects of computerized materials supporting construction of mental models, because the outcomes of this question for individual participants can assist in assigning the participants to study groups (one should either equalize groups on the MATHSE variable or take MATHSE as a covariate in the analysis).

Likewise, we can ask participants on their self-assessed ability of acquiring mental models (SAAMM) prior to the intervention. Having a single SAAMM question would have similar advantages to a single MATHSE question; provided the SAAMM question predicts learning outcomes, as concerns mental models acquisition.

The goal of this paper is to re-analyze data from our four studies (researching personalization principle (Mayer, 2009) in new contexts) from the perspective of the math self-efficacy—mental models learning link and the SAAMM—mental models learning link. We devised one SAAMM question and one MATHSE question (see Table 1), administered it (with other questions) prior to each intervention and correlated it with test scores (obtained after the intervention) and with several other variables.

In the next section, we briefly introduce the personalization principle we originally studied (in order to describe the context). Afterwards, we describe the interventions we used and details of our four studies. Finally, we present the results of the current analysis and its implications.

Table 1: Math self-efficacy and Self-assessed ability of acquiring mental models questions.

MATHSE and SAAMM questions		
Question	Text	Scale
Math self-efficacy (MATHSE)	"Check one of the following to indicate your knowledge of mathematics..."	1 – very small; 6 – very large; Likert scale
Self-assessed ability of acquiring mental models (SAAMM)	"Imagine you will be examined on the history of shipping traffic in the 19th century. A week before the exam, the examiner proposes you that you can learn just one of the following two things: a) the names of British steamboats from the second half of the 19th century, including their displacement and their propeller type, or b) how these steamboats' propellers work. There are over sixty of steamboats and five functionally-distinct types of propellers. What would you prefer to learn?"	1 – definitely (a); 7 – definitely (b); Likert scale

2. Personalization principle

Personalization principle is one of the well-known principles of multimedia learning (Mayer, 2009). These principles can be taken as guidelines for designers developing multimedia learning materials. Multimedia learning materials are defined, in this context, as computerized or non-computerized materials that combine texts and pictures; including books, animations, simulations and games for learning (cf. Clark & Mayer, 2011).

Personalization principle states that learners will learn better when textual instructions are in a conversational style compared to instructions given in a formal/neutral style. There are more possible explanations for this principle, including the increase of the learners' interest and the presence of social cues in the personalized instructions. A recent meta-analysis (Ginns, Martin, & Marsh, 2013) reported that the effect of personalization is robust and statistically reliable (Cohen's $d = 0.54$; as concerns deep understanding); however, certain boundary conditions remain to be investigated. These include personalization in languages other than English or longer treatments (longer than 35 min – such as complex simulations or games).

The original purpose of our research project was to investigate the personalization principle in the Czech language. To this end, we run four studies with college audience comparing learning effectiveness of computerized learning materials with personalized vs. non-personalized instructions (i.e., the between-subject design). At the same time, we investigated various supplementary research questions, one of them being if self-assessed ability of acquiring mental models and math self-efficacy predict test scores. The motivation behind this question is stated in Sec. 1. We could address this particular question easily because our participants' pools included students with high as well as low mathematical knowledge.

As already said, the purpose of this particular paper is to re-analyse data from our four experiments on personalization principle from the perspective of the hypothetical link between MATHSE/SAAMM and learning outcomes. In general, we found only small differences between our personalized and non-personalized conditions. Therefore, for present purposes, data from the two conditions were collapsed for each experiment.

3. Method

3.1 Participants

Participants were Czech college students (Mean age = 22.72; SD = 3.53) with intentionally diverse backgrounds (primarily computer science, psychology, and arts) and predominantly low prior knowledge of the topic of the respective intervention. They participated for course credit in one of four studies on personalization principle (N = 75, 85, 76, 41; for Experiment 1 – 4, respectively).

3.2 Interventions

Each participant interacted with one of the following computerized material: in Experiment 1 and 2, it was a 2-3 hours long computer simulation resembling a serious game on the topic of beer brewing; in Experiment 3, it was a few minutes long animation explaining the process of lightning formation; in Experiment 4, it was a comparable animation on functioning of biological wastewater treatment plant. In each experiment, each participant received either personalized or non-personalized version of the intervention.

3.2.1 Beer brewing simulation

The intervention for Experiments 1 and 2 was a 2-3 hours long computer simulation, in which the learner's goal was to acquire a mental model of the beer brewing process, as described in detail in (Brom et al., 2014). The simulation is for a single user, it is interactive and, as concerns its complexity, resembles a computer game with schematic graphic (Figure 1). The graphical user interface consists of the several elements, including: textual instructions, an animation panel showing the content of the fermentation vessel, a supplementary explanation panel relaying the meaning of graphical elements, panels with graphs and histograms showing the amount of ingredients in the product, an adjustable thermometer, and buttons for controlling the processes. The simulation also provides feedback via an "assessment" button. The non-personalized instructions have, in total, 6,750 words; the personalized version is slightly longer (the differences between the personalized and non-personalized instructions are unimportant for present purposes; they are detailed in Brom et al., 2014).

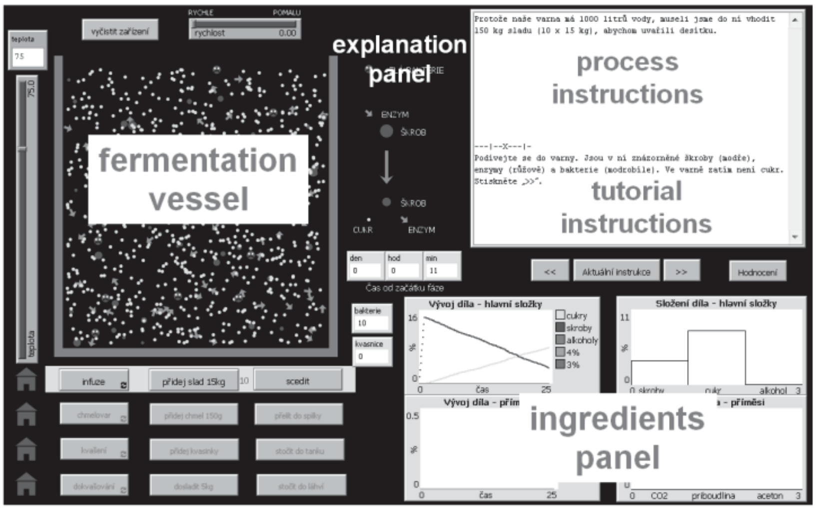

Figure 1: Screenshot of the beer brewing simulation used in Experiment 1 and 2. The main elements of the graphical interface are described. The instructions are in Czech. Note that the "ingredients panel" is composed of two graphs and two histograms). Also note that the interface is in colour in the application

The whole simulation has four parts. These parts include a) a tutorial, b) a segment that demonstrates in a linear fashion how to brew beer from beginning to end, when every step is done correctly, c) a segment that demonstrates the consequences of making errors of not following the standard procedure and d) a segment in which the learner has to brew his/her several beers of a specific type in the virtual brewery. The phenomenon being modelled – the process of beer brewing – is motivating enough for participants to stay with the simulation over the whole period (Brom et al., 2014).

The simulation, including textual instructions, was developed by the research team solely for research purposes.

3.2.2 Lightning formation animation

The intervention for Experiment 3 was a black-and-white, few minutes long animation explain the process of lightning formation. The experiment was a (so far unpublished) replication of the original experiment on personalization principle by Moreno & Mayer (2000; Exp. 1, 2) in Czech context. Therefore, the animation is as accurate replica of the original animation as possible[1] and the instructions are translated from the original animation.

The animation features 16 screens and each screen displays instructions at the bottom of the screen (Figure 2). The non-personalized instructions have, in total, 260 words; the personalized version is slightly longer. The user can only proceed forward – by clicking the "next" button (otherwise, the animation is non-interactive).

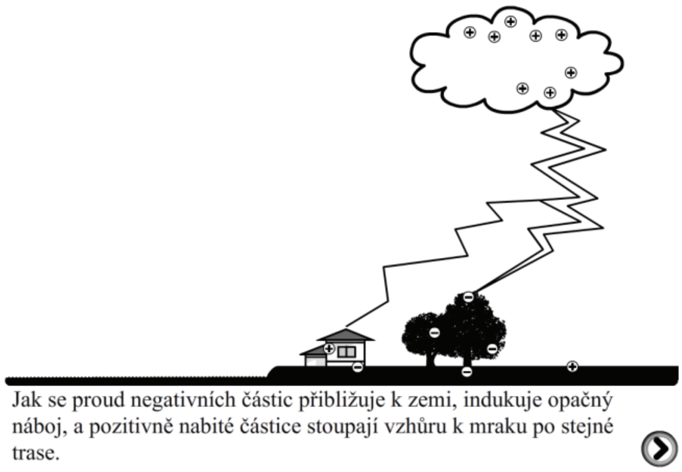

Figure 2: Screenshot of the lightning formation animation used in Experiment 3 (instructions in Czech). The "next" button is in the bottom right corner

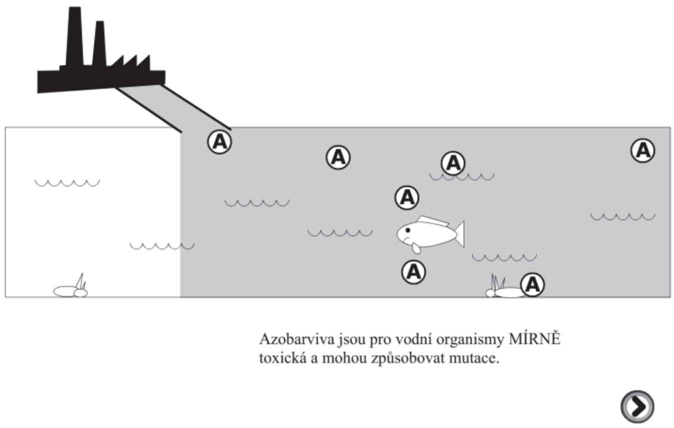

Figure 3: Screenshot of the biological wastewater treatment plant animation used in Experiment 4 (instructions in Czech). The "next" button is in the bottom right corner

3.2.3 Biological wastewater treatment plant animation

In Experiment 4, we attempted to replicate Experiment 3 with a different, yet highly comparable, animation. For this purpose, we created a black-and-white animation on functioning of biological wastewater treatment plant. This animation is, again, a few minutes long. It features 19 screens, each with instructions at the bottom of the screen (Figure 3). The non-personalized instructions have, in total, 300 words; the personalized version is slightly longer. The user can, again, interact with the animation only by clicking the "next" button.

[1] Discussed with the author of the original experiment (email from R. Mayer, 15th March, 2013).

3.3 Procedure and pen-and-paper materials

For each of the four experiments, the following applied. Prior to intervention, participants were administered self-assessment pre-tests with about seven questions of the following type: "Please, indicate your knowledge of X on the scale 1 – very small; 6 – very large." (see Brom et al., 2014 for the beer brewing pre-test; the pre-test for the lightning formation animation was from electro-physics, and it was from wastewater treatment and organic chemistry for the biological wastewater treatment animation – the latter two pre-tests were modelled according to Moreno & Mayer, 2000). The participants also received (among others) the following key questions: on math self-efficacy and self-assessed ability of acquiring mental models (denoted as MATHSE and SAAMM variables; see Table 1 for exact wording), and on computer game-playing frequency (denoted as Games variable; Scale 1 (less than 1 h a week) – 4 (more than 10 h a week)).

Afterwards, the participants interfaced with the intervention; and they proceeded at their own pace. Immediately after the intervention, the participants received knowledge post-tests. Every participant received (among others) a transfer test with four to seven (depending on the intervention) open-ended questions (Transfer variable). Transfer tests are particularly well suited for investigation of understanding of a process/model and they are gold standard in the field of multimedia learning (Mayer, 2009) (see Brom et al., 2014, for details concerning the beer brewing test, and Moreno & Mayer, 2000, for the details concerning the lightning formation test; the structure of the wastewater treatment plant test mirrored the lightning formation test). The tests were graded based on pre-prepared list of idea units; each idea unit was rewarded 1 point and partially correct idea units .25 or .5 points (exact wording was not required). The learners could obtain up to 32 (Exp. 1, 2), 20 (Exp. 3) or 25 (Exp. 4) points.

In the case of the beer brewing simulation (Exp. 1, 2), participants also returned to the laboratory one month later to complete the second round of knowledge tests, including the transfer test (see Brom et al., 2014 for details). Therefore, in the present analysis, transfer tests scores, for Experiments 1 and 2, are taken as the average of the scores from the immediate and delayed transfer tests. In the delayed testing session, participants also received a graphing test (Graphing variable; Scale 0 (nothing) – 9 (maximum)) and, in Experiment 2, also a brief test of 2D spatial visualization skills (Spatial variable; Scale 0 (nothing) – 7 (maximum)). The graphing test is a shortened version of the test of McKenzie & Padilla (1986) and the spatial visualization skills is a shortened version of the test of Slezakova & Molnar (2011).

3.4 Data analysis

Since the data from the two conditions were collapsed for present purposes, we did not use condition as an independent variable. We thus used the following linear regression model for explaining the transfer test scores:

$$Y_i\text{ (transfer)} = \beta_0 + \beta_1\, Pretest_i + \beta_2\, MATHSE_i + \beta_3\, SAAMM_i + \beta_4\, Games_i + \beta_5\, Graphing_i + \beta_6\, Spatial_i + \varepsilon_i,$$

where $Pretest_i$, $MATHSE_i$, $SAAMM_i$, $Games_i$, $Graphing_i$, and $Spatial_i$ are variables for i-th student, Y_i (transfer) is transfer test score for *i*-th student, and $\varepsilon_i \sim N(0, \sigma^2)$ denotes the random error of *i*-th student.

Correlations were expressed using Pearson's r, and effect sizes were classified according to Cohen (1988) as small (~ .1), medium (~ .3), and high (~ .5). In one case, we also used the two-sample t-test and explained the differences using Cohen's d. The effect sizes were classified as small (~ .2), medium (~ .5) and large (~ .8).

4. Results

Table 2 shows the raw data for all experiments and Tables 3 – 6 correlations for individual experiments. As concerns validation of the MATHSE and SAAMM questions, we see (Table 2) that there are large differences between computer science (mathematics/physics) students and other students. For MATHSE, Cohen's d ranged between 1.72 – 2.57 (all differences were significant with $p < .001$); for SAAMM, Cohen's d ranged between 0.55 – 1.24 (all differences were significant with $p < .05$). Correlations between these two variables were .32 - .40, i.e., in small to medium range, suggesting the existence of a common denominator.

Table 2: Raw data for individual questions (split by participants' background: "CS" stands for background in Computer Science and/or Mathematics and/or and Physics; "Oth" stands for "Others", including psychology, new media studies and art). Means and SDs are given. The variables (Pretest, MATHSE, SAAMM, Games, Graphing, and Spatial) are described in Section 3.3

Raw data from Exp. 1 - 4												
	Pretest[a]		MATHSE[b]		SAAMM[c]		Games[d]		Graphing[e]		Spatial[f]	
	CS	Oth	CS	Oth	CS	Oth	CS	Oth	CS	Oth	CS	Oth
Exp. 1	5.18 (2.75)	5.42 (3.40)	5.28 (0.53)	2.89 (1.11)	6.50 (0.78)	5.64 (1.90)	2.00 (0.91)	1.47 (0.69)	7.23 (2.13)	6.60 (2.13)	N.A.	N.A.
Exp. 2	4.53 (3.20)	5.36 (3.17)	5.31 (0.57)	3.59 (1.26)	6.38 (1.18)	5.28 (1.89)	2.08 (1.11)	1.37 (0.53)	6.31 (2.03)	5.22 (2.13)	5.08 (1.46)	4.53 (1.93)
Exp. 3	8.78 (4.60)	3.83 (2.98)	5.06 (0.80)	2.90 (1.15)	6.17 (1.38)	4.55 (2.25)	1.67 (0.97)	1.34 (0.76)	N.A.	N.A.	N.A.	N.A.
Exp. 4	4.67 (2.69)	2.54 (2.12)	5.31 (0.63)	3.21 (1.07)	6.77 (0.44)	4.54 (2.15)	2.15 (0.90)	1.11 (0.31)	N.A.	N.A.	N.A.	N.A.

Notes:
N.A. means that this inventory was not administered in this particular experiment.
[a]Scale: 0 (minimum) to 32 (Exp. 1, 2) (maximum) or 20 (Exp. 3) or 25 (Exp. 4).
[b]Scale: 1 (very small) to 6 (very large).
[c]Scale: 1 (definitely the names and propeller types) to 7 (definitely propellers' functioning).
[d]Scale: 1 (less than 1 h a week) to 4 (more than 10 h a week).
[e]Scale: 0 (nothing) to 9 (maximum).
[f]Scale: 0 (nothing) to 7 (maximum).

Table 3: Correlation matrix for Experiment 1. Numbers in parentheses denote the number of participants entering the respective correlation. "N.A." means that at least one of the respective tests was not administered. (†$p < .10$ *$p < .05$ **$p < .01$ ***$p < .001$)

Correlation matrix for Experiment 1						
	Transfer	Pretest	MATHSE	SAAMM	Games	Graphing
Transfer	-					
Pretest	.02 (69)					
MATHSE	.37**(69)	-.02 (73)				
SAAMM	.28*(70)	-.14 (74)	.32**(74)			
Games	.39***(70)	.10 (74)	.26*(74)	.28*(75)		
Graphing	.50***(70)	.01 (69)	.26*(69)	.24*(70)	.20† (70)	
Spatial	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.

Table 4: Correlation matrix for Experiment 2. Numbers in parentheses denote the number of participants entering the respective correlation. "N.A." means that at least one of the respective tests was not administered. (†$p < .10$ *$p < .05$ **$p < .01$ ***$p < .001$)

Correlation matrix for Experiment 2						
	Transfer	Pretest	MATHSE	SAAMM	Games	Graphing
Transfer	-					
Pretest	.24*(84)					
MATHSE	.30**(84)	-.13 (85)				
SAAMM	.40***(84)	-.03 (85)	.40***(85)			
Games	.41***(84)	.00 (85)	.24*(85)	.24*(85)		
Graphing	.43***(84)	.16 (84)	.36***(84)	.29**(84)	.26*(84)	
Spatial	.19† (83)	-.03 (83)	.16 (83)	.19† (83)	.14 (83)	.22*(83)

Table 5: Correlation matrix for Experiment 3. Numbers in parentheses denote the number of participants entering the respective correlation. "N.A." means that at least one of the respective tests was not administered. († $p < .10$ * $p < .05$ ** $p < .01$ *** $p < .001$)

Correlation matrix for Experiment 3						
	Transfer	Pretest	MATHSE	SAAMM	Games	Graphing
Transfer	-					
Pretest	.38***(76)					
MATHSE	.22† (76)	.60***(76)				
SAAMM	.24*(76)	.42***(76)	.38***(76)			
Games	.01 (76)	.22† (76)	.09 (76)	.15 (76)		
Graphing	N.A.	N.A.	N.A.	N.A.	N.A.	
Spatial	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.

Table 6: Correlation matrix for Experiment 4. Numbers in parentheses denote the number of participants entering the respective correlation. "N.A." means that at least one of the respective tests was not administered. († $p < .10$ * $p < .05$ ** $p < .01$ *** $p < .001$)

Correlation matrix for Experiment 4						
	Transfer	Pretest	MATHSE	SAAMM	Games	Graphing
Transfer	-					
Pretest	.39*(37)					
MATHSE	.57***(41)	.43**(37)				
SAAMM	.48**(41)	.34*(37)	.40**(41)			
Games	.48**(41)	.16 (37)	.47**(41)	.37*(41)		
Graphing	N.A.	N.A.	N.A.	N.A.	N.A.	
Spatial	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.

Table 7 shows the main results. As concerns Experiment 1 and 2 (the complex beer brewing simulation), we see that both MATHSE and SAAMM variables are indeed correlated to test scores with small to medium effect size (Tables 3, 4); however, these correlations can be largely explained by game playing frequency and graphing skills for Experiment 1 and 2, and also by pre-test score in Experiment 2 (Table 7). The only significant MATHSE/SAAMM regression coefficient is for the SAAMM question for Experiment 2 (Table 7). Because the simulation resembles a computer game and uses graphs, this outcome is in fact not surprising. Could we find a clearer effect of math self-efficacy and self-assessed ability of acquiring mental models on learning outcomes using non-game-like computerized materials with no graphs?

In Experiment 3 (the lightning formation animation), we found a small to medium effect size relationship between test performance and MATHSE/SAAMM variables (Table 5), but these two variables are also highly correlated to the pre-test score, which turned out to be the only significant predictor of the test performance (Table 7). This finding is also not that surprising because the animation's content was related to electro-physics and the link between the math self-efficacy and prior electro-physical knowledge (and achievement) is understandable. Could we find a clear effect of math self-efficacy and self-assessed ability of acquiring mental models on test performance with an animation unrelated to mathematics?

We used such animation in Experiment 4 (the biological wastewater treatment plant animation). Now, the test performance is highly correlated to the MATHSE variable (Table 6) and this variable is the only significant predictor of the test performance (Table 7). The SAAMM—test performance correlation is also high range (Table 6) and the SAAMM standardized regression coefficient is notable (0.26; Table 7) yet non-significant. This could be due to small power of the statistical test (N = 41 for Experiment 4).

Table 7: Standardized regression coefficients for each experiment. Coefficient of determination for each model is shown in the first column. The other columns show standardized regression coefficients (β) for each variable with 95% confidence intervals (CI). "N.A." means that the variable was not used in the model. (†p < .10 *p < .05 **p < .01 ***p < .001)

Main results – regression coefficients															
		Intercept		Pretest		MATHSE		SAAMM		Games		Graphing		Spatial	
Exp.	R^2	β	CI 95%	β	CI 95%	β	CI 95%	β	CI 95%	β	CI 95%	β	CI 95%	β	CI 95%
1	0.38	-0.04	[-0.25, 0.16]	0.05	[-0.17, 0.27]	0.2†	[-0.02, 0.41]	0.08	[-0.17, 0.34]	0.25*	[0.03, 0.46]	0.38***	[0.17, 0.59]	N.A.	N.A.
2	0.39	0.00	[-0.18, 0.18]	0.22*	[0.03, 0.41]	0.09	[-0.12, 0.30]	0.23*	[0.03, 0.44]	0.27**	[0.08, 0.46]	0.22*	[0.01, 0.43]	0.05	[-0.14, 0.24]
3	0.16	0.00	[-0.22, 0.22]	0.37*	[0.08, 0.65]	-0.04	[-0.31, 0.24]	0.11	[-0.13, 0.36]	-0.08	[-0.30, 0.15]	N.A.	N.A.	N.A.	N.A.
4	0.45	0.02	[-0.25, 0.30]	0.13	[-0.19, 0.44]	0.4*	[0.03, 0.77]	0.26	[-0.05, 0.56]	0.14	[-0.21, 0.50]	N.A.	N.A.	N.A.	N.A.

5. Discussion and conclusion

In all four studies, we found a relationship (i.e., a correlation) between math self-efficacy, assessed by our single question, and test performance, and also between self-assessed ability of acquiring mental models, assessed by a single question, and test performance (Tables 3 – 6). This relationship tended to be in medium to high ranges. Despite this relationship was masked and/or mediated by other variables in all treatments, except for the non-game-like animation whose content was unrelated to mathematics/physics (Table 7), its existence is a useful practical finding. The point is that in studies with between-subject design (using interventions supposed to teach mental models acquisition, such as simulations and games), the groups should be equalized with respect to these two variables. Otherwise, one group can outperform the other not due to the intervention's features but because participants from one group were predisposed to acquire the mental model easily. The equalization of groups can be done easily because researchers can administer these two questions in pre-tests and the outcome can immediately assist them in assigning participants to groups. Alternatively, the researchers can use these two variables as covariates. (Of course, it should be also tested if these variables indeed correlate to test scores in the experiment in question.)

Our finding could have also theoretical implications, if thoroughly replicated. This implication would be that people experienced in manipulating mathematical objects in working memory may be in advantage as concerns general mental models acquisition. This ability would probably be related to increased spatial abilities of these learners (cf. Ackerman et al., 1995; Höffler, 2010).

As concerns limitations of the present study, validating the MATHSE/SAAMM questions against real mathematical tests' outcomes would be vital. Note also that in Experiment 4, we did not administer graphing test, which is a drawback from the perspective of the present study (this test was not originally planned for Experiment 4). As for the test of spatial abilities: even though we have not found the link between spatial abilities and mathematical self-efficacy (Table 4), larger tests of spatial abilities can indicate differently (cf. Ackerman et al., 1995).

There are also open issues. First, it is unknown when the influence of frequent gaming on transfer test scores is mediated by mathematical skills (due to math skills—frequent gaming relationship) and when it is causal (due to materials' game-like appearance). Second, boundary conditions of this effect must be investigated: for instance, our different study indicated that our math self-efficacy question is unrelated to test performance for high school audience. We believe that the reason is that college students (who knew that diverse students participated in the experiment) related this question to general population whereas high-school students to their class (i.e., school grade), brining substantial noise due to inter-class differences.

Considering all points together, our tentative suggestion is that researchers investigating acquisition of mental models by computerized materials (using between-subject design with quantitative measures) should consider equalizing groups with respect to participants' mathematical ability. The level of this ability can be probably indicated indirectly by our two questions on mathematical self-efficacy and self-assessed ability of acquiring mental models. In future, rigorous validation of these questions and investigation of boundary conditions would be useful.

Acknowledgements

This research was partially funded by the research project nr. 15-14715S supported by the Czech Grant Science Foundation (GA ČR) (for C. B.) and by the student grant GA UK no. 68413 (for F. D.). We are especially thankful to Richard Mayer for commenting on early versions of the lightning formation animation, Edita Bromová for developing the interventions, and for Martin Pergel and Kateřina Svobodová for helping with design of the beer brewing simulation and the biological wastewater treatment plant animation. We also thank the staff of Laboratory of Behavioral and Linguistic studies who helped us with management of the subject pool and provided places for part of the experiment. Last but not least, we thank all research assistants who helped to conduct the experiments, most notably: Tereza Selmbacherová, Tereza Stárková, Michaela Buchtová, Viktor Dobrovolný, Ondřej Smíšek, Martina Denemarková, Martina Stejskalová, and Nela Bendová. The human data were collected with APA ethical principles in mind.

References

Ackerman, P. L., Kanfer, R. and Goff, M. (1995) "Cognitive and noncognitive determinants and consequences of complex skill acquisition", *Journal of Experimental Psychology: Applied*, Vol. 1, No. 4, pp 270-304.

Brom, C., Bromova, E., Dechterenko, F., Buchtova, M., Pergel, M. (2014) "Personalized Messages in a Brewery Educational Simulation: Is the Personalization Principle Less Robust than Previously Thought?", *Computers and Education*, Vol. 72, pp. 339-366.

Clark, R. C. and Mayer, R. E. (2011) *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*, John Wiley & Sons.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.), New Jersey: Lawrence Erlbaum.

Ginns, P., Martin, A. J. and Marsh, H. W. (2013) "Designing instructional text in a conversational style: A meta-analysis", *Educational Psychology Review*, Vol. 25, No. 4, pp 445-472.

Hackett, G. and Betz, G. (1989) "An exploration of the mathematics self-efficacy/mathematics performance correspondence", *Journal for Research in Mathematics Education*, Vol. 20, No. 3, pp 261-273.

Höffler, T. N. (2010) "Spatial ability: Its influence on learning with visualizations—a meta-analytic review", *Educational psychology review*, Vol. 22, No. 3, pp 245-269.

Mayer, R. (2009) *Multimedia Learning* (2nd ed.), Cambridge University Press.

McKenize, D. L., Padilla, M. J. (1986) "The construction and validation of the test of graphing in science (TOGS)", *Journal of Research in Science Teaching*, Vol. 23, No. 7, pp 571 - 579.

Moreno, R., & Mayer, R. E. (2000) "Engaging students in active learning: The case for personalized multimedia messages", *Journal of Educational Psychology*, Vol. 92, pp 727– 733.

Papert, S. (1980) *Mindstorms: Children, computers, and powerful ideas*, Basic Books, Inc.

Slezakova, J., Molnar, J. (2011) "Testování geometrické představivosti" ["Testing geometrical imagery"], *E-Pedagogikum*, Vol. 3/2011, pp 80-96. [in Czech]

Wouters, P., van Nimwegen, C., van Oostendorp, H. and van der Spek, E. D. (2013) "A Meta-Analysis of the Cognitive and Motivational Effects of Serious Games", *Journal of Educational Psychology*, Vol. 105, No. 2, pp 249-265.