Representations of equation-based models are not created equal

Alexander Pollok
Institute of System Dynamics and Control
DLR German Aerospace Center
Oberpfaffenhofen, Germany
alexander.pollok@dlr.de

Andreas Klöckner
Institute of System Dynamics and Control
DLR German Aerospace Center
Oberpfaffenhofen, Germany
andreas.kloeckner@dlr.de

Dirk Zimmer
Institute of System Dynamics and Control
DLR German Aerospace Center
Oberpfaffenhofen, Germany
dirk.zimmer@dlr.de

ABSTRACT

For equation-based modelling languages, modelling experts have many degrees of freedom when building a model from scratch. One of the most basic choices the expert faces is the mode of representation. The same system can be represented for instance as a block-diagram, by writing down the physical equations, by writing an algorithm, or by graphically connecting ready-made subcomponents. To give some guidance in this aspect, an experiment was conducted to measure the effects of different representations on various tasks. Participants had to identify models and predict their transient response. Both the time to execute the task and the correctness of the answer were measured. Participants also had to rate their confidence regarding the models. Results showed that tasks were executed much faster for graphical representations than for block-digrams. Equation-based and algorithm-based models can be grouped in the middle. The same results hold for rated confidence. Interestingly, the amount of errors was similar for all representations. Apparently, modelling experts largely compensate for difficulty by taking their time.

CCS CONCEPTS

• **Computing methodologies** → **Modeling methodologies**; **Simulation languages**; • **Applied computing** → *Engineering*;

KEYWORDS

modelling, programming aids, coding style, psychology, software ergonomics, trial, experiment

ACM Reference format:
Alexander Pollok, Andreas Klöckner, and Dirk Zimmer. 2017. Representations of equation-based models are not created equal. In *Proceedings of 8th International Workshop on Equation-Based Object-Oriented Languages and Tools, Wessling, Germany, December 1, 2017 (EOOLT'17)*, 6 pages.
https://doi.org/10.1145/3158191.3158200

1 INTRODUCTION

Mental representations of computer programs are different for various programming language, and programming performance is higher if there is a match between the external representation and the mental representation ([4], [6], [3], [11]). We postulate that a similar dependency exists in the field of physical modelling.

In equation-based languages, several different representations are common. This is also true within the scope of individual equation-based languages. For example, in Modelica it is possible to model a system by writing down the necessary physical equations, or alternatively by joining together the necessary subcomponents graphically.

In this paper, an experiment is conducted to find differences in cognition between different representations of equation-based simulation models. The paper is structured as follows: Section 2 explains the experiment in detail. This Section is further structured into descriptions of the experiment design, the participants, the materials used, the experiment procedure and the internal review process. In Section 3, the results of the experiment are shown and analysed. The results are discussion in Section 4. Section 5 concludes the paper.

We expect that the typical reader of this paper has a background from engineering, computer science or similar. However, many concepts of this paper are taken from psychology and statistics instead. To keep the experience enjoyable to the readers, these concepts are explained in footnotes in a somewhat uncouth and pragmatic fashion.

2 METHOD

2.1 Design

Participants were shown 4 different Modelica models, or representations, of physical systems. They were asked to *identify* those systems, *predict* the transient response, and *rate* their confidence for those models. The models were created in such a way, that they represented each physical system in four different ways. The time needed by the participants for the identification and prediction tasks was measured.

For each of the 4 physical systems considered, 4 different Modelica models were created, to a total of 16 different models. Participants were without their knowledge split up into random groups of equal size. Each group was presented with one representation of each physical system. To eliminate first order training and carryover effects, a balanced latin square design[1] as described in [12]

EOOLT'17, December 1, 2017, Wessling, Germany

ACM ISBN 978-1-4503-6373-0/17/12…$15.00
https://doi.org/10.1145/3158191.3158200

[1] If participants have to do multiple tasks, one after another, several effects are at work. One task might be more difficult than another, resulting in participants taking longer to complete this task. On the other hand, if the tasks are similar, participants will have some experience by the time they get to the last task, resulting in a shorter time to complete the task. Also, some of the tasks might be a good preparation for others, and vice versa. To isolate this effects, a balanced latin square design can be used. Here, the participants are divided into several groups. The groups are assigned to tasks in a specific order, designed to eliminate the experience (or training) and preparation (or carryover) effects under the assumptions that these effects are first order.

was used to assign representations to physical systems for each group of participants. Effects of physical systems are not considered in this work, therefore the order of physical systems was the same for each group. The effect is thus indistinguishable from training effects and it is accounted for by the balanced latin square design.

The expectation was that the percentage of correctly executed tasks, as well as timings and reported confidence would vary between the different representations.

2.2 Participants

Since knowledge in Modelica was necessary for participation in the experiment, participants were recruited from the Modelica User Groups Sachsen, North America, Japan, Hamburg and Baden-Württemberg, as well as colleagues of the authors at the Institute of System Dynamics and Control. Participants were not compensated. Over the course of one month, 98 participants took part in the experiment. From those 20 participants were excluded based on having not completed the tasks (5), unrealistic timings and reported technical problems (5), reports of being interrupted (9), or reporting a Modelica experience of zero (1). Among the 78 remaining participants, 8 reported being female, 69 male, and 1 other. They were aged between 23 and 63 years (mean $M = 36.7$, standard deviation $SD = 8.7$). They had previous experience with Modelica between 0.1 and 17 years (mean $M = 5.2$, standard deviation $SD = 4.4$). Participants reported their professional background based on the provided categories shown in Table 1. Notably, most participants were engineers and there were no computer scientists. Table 1 shows the place of living of the participants ordered by continents. Most of them lived in Europe or North America.

Mechanical Engineering or similar	43	Europe	51
Electrical Engineering or similar	11	North-America	14
Mathematics	9	Asia	12
Engineering (other)	8	(not answered)	1
Physics	3		
Other	2		
Science (other)	2		
Computer science	0		
(not answered)	1		

Table 1: Origin and Background of participants

2.3 Material

For four different physical systems, four models were created each in Modelica, resulting in 16 models in total. The types of systems and representations are listed in Table 2.

Models were developed in such a way that they were behaving equally for each system, while keeping the models as simple as possible. All models for the Spring-Damper system are shown in Figures 1, 2, 3 and 4.

For each of the systems, eight names were created. Of those eight names, four were physically motivated, four were motivated from a system dynamics point of view. From each group of four, one of them correctly described the system. For example, the Spring-damper system was assigned the following names (correct names written in bold):

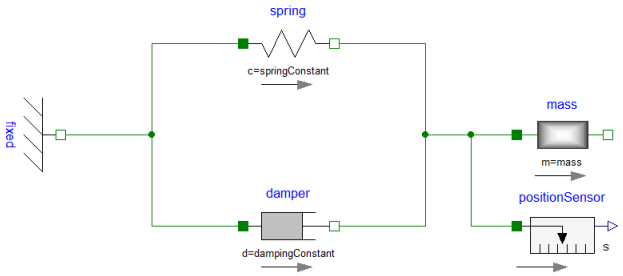

Figure 1: MSL - graphical model of a Spring-Damper

```
model Test1
  parameter Modelica.SIunits.Length          position0;
  parameter Modelica.SIunits.DampingConstant dampingConstant;
  parameter Modelica.SIunits.SpringConstant  springConstant;
  parameter Modelica.SIunits.Mass            mass;

  Modelica.SIunits.Length       position;
  Modelica.SIunits.Velocity     velocity;
  Modelica.SIunits.Acceleration acceleration;
equation
  //force balance
  0 = -    acceleration      * mass
      -      velocity        * dampingConstant
      - (position-position0) * springConstant;
  der(velocity) = acceleration;
  der(position) = velocity;
end Test1;
```

Figure 2: EQ - equation-based model of a Spring-Damper

```
model Test2
  parameter Modelica.SIunits.Length          position0;
  parameter Modelica.SIunits.DampingConstant dampingConstant;
  parameter Modelica.SIunits.SpringConstant  springConstant;
  parameter Modelica.SIunits.Mass            mass;

  Modelica.SIunits.Length   position;
  Modelica.SIunits.Velocity velocity;
  Modelica.SIunits.Force    force;
algorithm
  force          := -      velocity       *dampingConstant
                    - (position-position0)*springConstant;
  der(velocity) := force/mass;
  der(position) := velocity;
end Test2;
```

Figure 3: ALG - algorithm-based model of a Spring-Damper

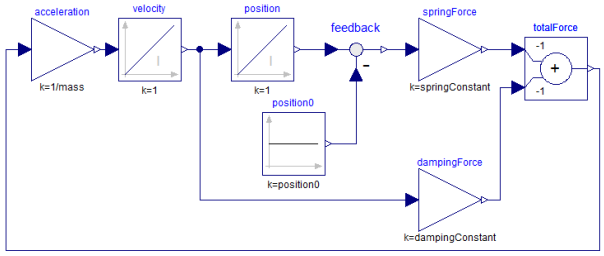

Figure 4: BLOCK - block-diagram model of a Spring-Damper

systems		representations	
SD	a parallel spring-damper with connected mass	MSL	graphical models built from MSL components
OC	an electric resonant circuit without damping	EQ	models based on physical equations
T	two bodies in thermal contact	BLOCK	block diagrams built from the MSL.Block-package
BB	a mass bouncing on a floor	ALG	models implemented as algorithms

Table 2: Physical systems and representations used in the experiment

(1) **parallel spring-damper with connected mass** (2) force amplifying system with damping (3) electric resonant circuit without damping (4) operational amplifier (5) **second-order system with damping** (6) first-order system with connected integrator (7) proportional, integral, and derivative systems connected in parallel (8) non-linear second-order system without damping.

Furthermore, for each of the systems, nine images were created that illustrated possible trajectories. Of those trajectories, only one had an adequate connection to the system. Figure 5 shows the possible answers for the spring-damper system (correct answer framed).

2.4 Procedure

The experiment was conducted online using SoSci survey [7]. Tasks were implemented as multiple-choice tests. Timings were measured using asynchronous java-script and XML with an accuracy of a few milliseconds, independently of the status of the internet connectivity of the participant.

At the start of the experiment, participants were shown a starting page, and were without their knowledge assigned to one of four groups. The assignment to one of these groups was based on chance and on the assignment of prior participants. This was done to ensure that the number of data points for all groups was roughly equal. Before each task, participants were given a short introduction to the task. For the first two tasks, this introduction also included a request to *"hurry up"*. This was done to remind participants that the time to execute the task is important and being measured. For the third task, rating of confidence, timing was not a concern. During the first task of the experiment, participants were shown four different combinations of systems and experiments and had to identify the systems in questions. The available answers for each combination were scrambled randomly. The assignment of the 16 combinations of systems and representations was done using a balanced latin square approach in the representations and constructed as shown in [1]. The assignment can be seen in Table 3.

During the second task, participants were shown the same four combinations. This time, they were asked to predict the transient response, or future dynamical behavior, of the system. Again, the order of answers were scrambled.

During the third task, participants were again shown the same four combinations in the same order. They were asked how much they agree with the following statement: *"I would feel confident using this model."* Participants could choose from a five-point Likert scale[2], ranging from *"-2 (Strongly disagree)"* to *"+2 (Strongly agree)*. This time, the order of answers was not scrambled.

	subtask 1	subtask 2	subtask 3	subtask 4
group 1	OC-BLOCK	T-EQ	SD-MSL	BB-ALG
group 2	OC-EQ	T-ALG	SD-BLOCK	BB-MSL
group 3	OC-ALG	T-MSL	SD-EQ	BB-BLOCK
group 4	OC-MSL	T-BLOCK	SD-ALG	BB-EQ

Table 3: Assignment of system/representation-combinations to groups

For each task, the first subtask doesn't give valid timing results, since the browser has to load the content first, and the participant has to understand the style of question. Therefore, for each task, an additional training subtask was added at the beginning. For example, at the beginning of the identification task, participants were shown a picture of bananas, together with the following question: *"Let us start with a small warmup question to get into the mood: What system is represented here (click one of the 2 correct options)?"*. Possible answers were: *"bananas"*, *"apples"*, *"yellow fruits"*, *"green fruits"*.

To lower the effects of stereotype threat[3] on the results of this paper, the actual tasks were executed by the participants at the beginning of the experiment. Only after conducting all three tasks (identification, prediction of transient response, rating confidence), personal information about the participants was gathered.

Participants were asked if they were interrupted during the experiment. They were also asked to estimate their Modelica experience in years, their professional background (dropdown-menu),

[2] A Likert scale is a psychometric scale to quantify responses. If participants are asked: "How much do you love your dog?", some answers might be: "a lot" or "with all my heart" or "meh...". Mapping these answers to a love-metric will be highly arbitrary at best. Here, Likert scales should be used instead. The better question will be: "How much do you agree with the following statement: I love my dog.", while participants can choose from three, five or seven (fun fact: the human brain cannot meaningfully differentiate between more than seven grades of a single concept) predefined answers, ranging from "agree" to "disagree". Odd numbers are used, because a neutral response has to be included. These answers are then translated to numbers. Often, equal distances between two neighboring answers are assumed. These numbers constitute the new attribute, which is far easier to analyse.

[3] Performance in any kind of task can be effected by so called stereotype threat: if participants are under the impression that a sociocultural group they belong to is negatively stereotyped for that particular task, performance gets distorted to fit that stereotype [2, 8, 10]. For instance, women underperformed men in a math test, when the test got described as producing gender differences; the difference could be eliminated when the test got described as not producing gender-difference [9].

Figure 5: Answer options for the transient responses of the damped spring-damper system (correct answer framed)

age, gender and place of living (dropdown-menu with continents). Finally, they could make a guess regarding the goal of the experiment and could give miscellaneous remarks. Interested participants could also leave their E-Mail address to be notified of the experiments results, those addresses were stored independently, for the results to remain anonymous.

2.5 Internal review

Models were individually reviewed and modified by Alexander Pollok, Andreas Klöckner and two additional colleagues at the Institute of System Dynamics and Control, to keep code and model quality equally high for each of the 16 models. Both system names and trajectories were checked for ambiguity by the same people. In some cases the problem definition was augmented to ensure that each problem had a unique solution, for instance by adding information about the sign of a damping constant. The complete experiment was run as a pretest by eight colleagues of the authors to ensure functionality and comprehensibility[4]. The data from those experiments was discarded.

3 RESULTS

The basic answer characteristics for the three main tasks are shown in Table 4. There are no significant[5] differences in error rates of identification ($\chi^2(3) = 0.69, p = 0.875$)[6] nor prediction ($\chi^2(3) = 1.74, p = 0.627$). The type of answer (mathematical or physical) in the identification task has significant differences over the four representations ($\chi^2(3) = 9.64, p = 0.022$). The p-values of the pairwise differences are assessed in Table 5a. In summary, the MSL representation leads to significantly more physical answers as compared to representation as equation (EQ). Table 4 also shows the mean rating estimates and standard errors for each representation. Models are rated best, if they are built with MSL, and worst, if they are built as block diagram. Details on these values are described later in this section.

The influence of the type of answer in the identification task on the other measures is shown in Table 5b. Holm correction[5] is used[7]. There are no significant effects of the answer type on identification correctness ($\chi^2(1) = 3.80, p = 0.051$) nor prediction correctness ($\chi^2(1) = 0.24, p = 0.627$). Estimated rating means are included for illustration. There is no significant effect on these either.

The reaction times and ratings are subjected to a mixed-effects regression with random intercepts

$$time = a + b(participant) + b(representation) + b(id\_type) + b(id\_correct) \quad (1)$$

where a represents a fixed mean intercept, $b(participant)$ represents an additional intercept for each participant, and $b(representation)$, $b(id\_type)$ as well as $b(id\_correct)$ represent fixed effects of representation, type of answer in the identification task as well as correctness of the identification task. Table 6 shows the respective significance tests. There are significant effects for all fixed intercepts, for the influence of representation on identification reaction time ($F(3, 202) = 12.8, p < 0.001$)[8] and on rating ($F(3, 202) = 15.2, p < 0.001$).

There is no significant difference for representation forms in the prediction task ($F(3, 188) = 1.1, p = 0.343$). This could be due to the same model being presented twice, such that the reaction time is not influenced anymore by the representation of the model, but rather by the classification result from the prior task. This is supported by the finding that the type of answer from the naming task has indeed a significant influence on the timing for the prediction task ($F(1, 188) = 30.3, p < 0.001$). A physical answer type leads to a 8.8 s quicker response than a mathematical answer type ($t(188) = 5.11, p < 0.001$). In addition, a correct answer in the identification task carries over to the reaction time in the prediction task ($F(1, 188) = 5.5, p = 0.020$), indicating a more confident reaction in the second task for correct answers. This is confirmed by a significant effect from identification correctness on the actual rating ($F(1, 202) = 8.3, p = 0.004$).

Estimated means of the fitted models per representation type are shown in Table 6 along with their 95% confidence interval and letter displays for non-significant differences[9]. They are computed using the fitted values and averaging over participants, the different types of identification answer, and the correctness of the identification answer. Table 7 shows the values of these differences along with the respective test statistics.

The graphical MSL representation performs best in both the identification and in the rating task. The BLOCK diagram representation performs worst. There are no significant differences between

[4] It is considered good scientific practice to also mention any kind of reward or compensation offered to the participants. Therefore: Participants involved in pretests and internal reviews were compensated generously with cookies.

[5] Statistical significance is measured using the p-value. p is defined as the probability of obtaining even more extreme results than actually obtained, under the assumption that the null-hypothesis is true. A low value of p is therefore connected to a low probability of the found effects being produced by noise. $p = 1$ means that the hypothesis is not in any way supported by the data, while a purely theoretical $p = 0$ would indicate perfect agreement. Typically in social sciences, $p < 0.05$ is demanded for results to be considered significant. Even then, of 20 reported significant effects with $p = 0.05$ to be found, one of them will on average be invalid.

[6] $\chi^2(4-1)$ is a metric to test the statistical independence between two attributes. Here, the independence of the error rate and the 4 representation is tested. The smaller the number, the more probable it is that both attributes are independent.

[7] If multiple hypothesis are tested against a given significance level, the risk of having at least one false positive is automatically increased.
For a layman example, see http://xkcd.com/882. Holm correction is used to counteract this effect by adjusting the p-values.

[8] $F(4-1, 203-1)$ is a measure for the variance of the 203 results between the 4 groups, compared to the variance of the results inside of the groups. If this metric is high, the variance between the groups is comparatively high. This implies, that the groups are a good predictor.

[9] Letter displays indicate groups, in which no significant differences between the representation forms are found. For example, there is no significant difference in the identification reaction time between MSL representation and EQ representation. For this reason, they both contain the letter 1. There is a significant difference between MSL representation and ALG representation. For this reason, they do not have a letter in common.

	identification type		identification errors		prediction errors		rating	
	Math.	Physical	Correct	Wrong	Correct	Wrong	Mean	SE
MSL	10	68	59	13	61	13	0.87	0.16
EQ	25	53	54	9	51	14	0.43	0.16
ALG	19	59	64	10	61	11	0.27	0.15
BLOCK	24	54	65	11	57	17	−0.31	0.15

Table 4: Summary of answers grouped by type of representation

	MSL	EQ	ALG
EQ	0.043	—	—
ALG	0.399	1	—
BLOCK	0.058	1	1

(a) p values

	identification		prediction		rating	
	Correct	Wrong	Correct	Wrong	Mean	SE
Mathematical	77.3 %	22.7 %	78.1 %	21.9 %	0.37	0.11
Physical	87.6 %	12.4 %	81.6 %	18.4 %	0.26	0.15

(b) summary of answers grouped by type of identification

Table 5: Post-hoc tests for differences in the type of identification answer between representations with Holm correction[5]

		identification			prediction			rating		
	df	res	F	p	res	F	p	res	F	p
(Intercept)	1	202	599.4	0.000	188	462.3	0.000	202	44.3	<0.001
repr.	3	202	12.8	<0.001	188	1.1	0.343	202	15.2	<0.001
id_type	1	202	3.4	0.068	188	30.3	<0.001	202	1.2	0.284
id_correct	1	202	1.4	0.238	188	5.5	0.020	202	8.3	0.004

Table 6: Test statistics of linear regression models for the identification, prediction, and rating tasks

	reaction time for naming					rating of preference				
contrast	estimate	SE	df	t	p	estimate	SE	df	t	p
MSL - EQ	−4.58	3.10	202	−1.47	0.455	0.44	0.19	202	2.30	0.101
MSL - ALG	−11.75	2.90	202	−4.05	<0.001	0.59	0.18	202	3.30	0.006
MSL - BLOCK	−15.88	2.91	202	−5.46	<0.001	1.18	0.18	202	6.58	<0.001
EQ - ALG	−7.17	3.01	202	−2.38	0.084	0.15	0.19	202	0.83	0.842
EQ - BLOCK	−11.30	2.99	202	−3.79	0.001	0.74	0.18	202	4.03	<0.001
ALG - BLOCK	−4.14	2.84	202	−1.46	0.466	0.59	0.18	202	3.34	0.005

Table 7: Predicted differences between forms of representation for reaction times and rating

the EQ equation and the ALG algorithm representations. This can likely be explained by the two representations resembling each other strongly.

4 DISCUSSION

Modern modelling languages give modelling experts a large amount of freedom regarding the choice on how to represent a physical system. It was found that this choice has a significant impact on the performance of modelling experts for varying tasks. Based on the conducted experiments, graphical representations using the MSL should be preferred in most cases. Tasks were performed faster, and participants showed a higher amount of confidence. The opposite can be said for the representation by block diagrams. Achievable time savings can be rather important. A difference of up to 15.88 s was found between MSL and BLOCK representations. This corresponds to roughly one third of the absolute reaction time. Maximal rating difference was 1.18 between the same two representations, which also corresponds to one third of the total scale used for rating the model. Textual models, represented by equations or by algorithms, can be grouped between graphical MSL representations and block diagrams. No significant difference was found between these textual models. While equation-based representations might be more common in the Modelica community, the final choice will usually depend on the use case. For example, algorithms might

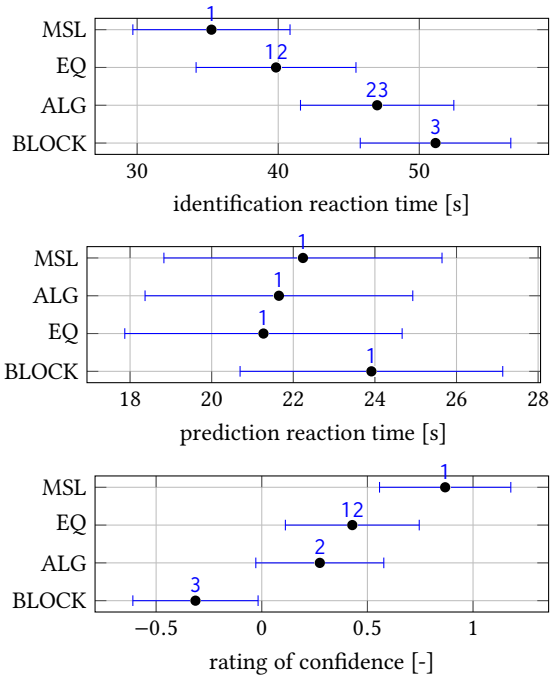

Figure 6: Predicted reaction times and rating with 95%-confidence intervals and letter displays for non-significant differences.

be used more often when controllers are modelled, or when non-physical computations are represented. Interestingly, the amount of errors was similar for all representations, while the necessary time was significantly different. This suggests that modelling experts completely compensate for tasks of higher difficulty by just taking more time. This is consistent with general finding on the speed-accuracy tradeoff and also manifests in the amount of confidence expressed explicitly.

While no direct influence of type of representation was found on the reaction time for the prediction task, there were significant influences mediated by the mental representation of the model. This can be counted as evidence towards more subconscious confidence in a physical than in a mathematical mental model. Another mediator variable is the correctness of the identification answer, which has significant influences not only on subconscious measure of prediction time, but also on the consciously expressed confidence in the model. Together, these results also suggest to prefer physical MSL models over equation-based or block-based models, in order to increase the confidence of users in the model and to increase the ease of working with these models.

4.1 External Validity

Results of these experiments are restricted to engineers and other users of equation-based modelling languages. This is due to the recruiting of the participants from the Modelica user groups. However, from an industry standpoint, this is also the group of people where such results are useful. While the individual timings of the experiment will surely be dependent on the background of the participants, the study design should compensate any such effects for the derived results. Even if a mechanical engineer will be faster at recognizing a mechanical spring-damper system compared to a computer scientist, this will be true for any of the different representations, which are compared relatively to each other. Results hold mainly for the western culture. Although a few asian participants took part in the experiment, there is no reasonable evaluation as to what differences exist between these cultural groups. However, no such differences would be expected anyway. For reasons of practicality, the used models were quite small. Due to the lack of similar studies, not much is known about the effects of model size and complexity. It is completely plausible that the best representation is dependent on the size and complexity of the system. Further studies are necessary to make generalizations in that direction.

Many other typical threats to external validity should not be relevant for this study. While participants might show increased performance due to the awareness of being observed (Hawthorne effect), this increase will be in effect for all of the tasks. Since only relative statements are derived from the resulting data, the Hawthorne effect should not decrease the external validity of this study. Similar arguments can be made for pre- and post-test effects, situational specifics, or Rosenthal effects (higher expectations lead to increased performance). Second order effects can not be ruled out, but their influence is expected to be small.

5 CONCLUSION

Modelling experts seem to understand different representations of equation-based simulation models faster than others. Based on experimental results, graphical representations are understood fastest, and representations based on block diagrams are understood slowest. Equation- and algorithm-based representations are somewhere in the middle. Similar results appear for ratings of confidence regarding the different representation types.

REFERENCES

[1] James V Bradley. 1958. Complete counterbalancing of immediate sequential effects in a Latin square design. *J. Amer. Statist. Assoc.* 53, 282 (1958), 525–528.
[2] Ap Dijksterhuis and Ad Van Knippenberg. 1998. The relation between perception and behavior, or how to win a game of trivial pursuit. *Journal of personality and social psychology* 74, 4 (1998), 865.
[3] Thomas RG Green and R Navarro. 1995. Programming plans, imagery, and visual programming. In *Human Computer Interaction*. Springer, 139–144.
[4] Thomas R. G. Green and Marian Petre. 1996. Usability analysis of visual programming environments: a cognitive dimensions framework. *Journal of Visual Languages and Computing* 7, 2 (1996), 131–174.
[5] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* (1979), 65–70.
[6] Jill H Larkin and Herbert A Simon. 1987. Why a diagram is (sometimes) worth ten thousand words. *Cognitive science* 11, 1 (1987), 65–100.
[7] Dominik J Leiner. 2014. SoSci survey (version 2.5.00-i)[computer software]. (2014).
[8] Margaret Shih, Todd L Pittinsky, and Nalini Ambady. 1999. Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological science* 10, 1 (1999), 80–83.
[9] Steven J Spencer, Claude M Steele, and Diane M Quinn. 1999. Stereotype threat and women's math performance. *Journal of experimental social psychology* 35, 1 (1999), 4–28.
[10] Claude M Steele. 1997. A threat in the air: How stereotypes shape intellectual identity and performance. *American psychologist* 52, 6 (1997), 613.
[11] Iris Vessey and Dennis Galletta. 1991. Cognitive fit: An empirical study of information acquisition. *Information systems research* 2, 1 (1991), 63–84.
[12] EJ Williams. 1949. Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Chemistry* 2, 2 (1949), 149–168.