Making the Case for Measuring Mental Effort*

Stefan Zugal
University of Innsbruck
Technikerstraße 21a
6020 Innsbruck, Austria
stefan.zugal@uibk.ac.at

Jakob Pinggera
University of Innsbruck
Technikerstraße 21a
6020 Innsbruck, Austria
jakob.pinggera@uibk.ac.at

Hajo Reijers
Eindhoven University of Technology
PO Box 513
NL-5600 MB Eindhoven, The Netherlands
h.a.reijers@tue.nl

Manfred Reichert
Universität Ulm
Building O27,
James-Franck-Ring
89069 Ulm, Germany
manfred.reichert@uni-ulm.de

Barbara Weber
University of Innsbruck
Technikerstraße 21a
6020 Innsbruck, Austria
barbara.weber@uibk.ac.at

ABSTRACT

To empirically investigate conceptual modeling languages, subjects are typically confronted with experimental tasks, such as the creation, modification or understanding of conceptual models. Thereby, accuracy, i.e., the amount of correctly performed tasks divided by the number of total tasks, is usually used to assess performance. Even though accuracy is widely adopted, it is connected to two often overlooked problems. First, accuracy is a rather insensitive measure. Second, for tasks of low complexity, the measurement of accuracy may be distorted by peculiarities of the human mind. In order to tackle these problems, we propose to additionally assess the subject's mental effort, i.e., the mental resources required to perform a task. In particular, we show how aforementioned problems connected to accuracy can be resolved, that mental effort is a valid measure of performance and how mental effort can easily be assessed in empirical research.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Experimental design

General Terms

Experimentation, Human Factors, Measurement

1. INTRODUCTION

Over the years, numerous conceptual modeling languages and associated modeling tools have been proposed [15]. In order to allow for an objective discrimination whether new proposals come along with improvements, the adoption of empirical software engineering has been advocated [4, 26]. Certainly, empirical research has been shown to be suitable for putting discussions on an objective basis. Still, in order to contribute to truly objective results, a valid experimental setup, as well as valid measurement methods are indispensable—slightest changes in the design might lead to fundamentally different outcomes [12].

In this work, we focus on empirical research that involves human activities, such as the creation, modification and understanding of conceptual models. Therein, various methods have been applied to identify differences. In particular, researchers have used modeling tasks [20], modification tasks [7] and sets of questions [5] to assess performance of conceptual modeling languages. In order to *measure* the outcome of tasks, typically *accuracy* and *duration* are analyzed (cf. [5, 7, 11, 20, 28]). Accuracy thereby refers to the percentage of correctly performed tasks, whereas duration refers to how long a subject required to perform a task. Even though accuracy is well-established and widely adopted, it is connected to two often overlooked problems. First, in order to identify differences with respect to accuracy, subjects need to commit errors. Hence, subtle differences that may be relevant, but do not lead to errors, cannot be identified (e.g., slight improvement of comprehensibility). Second, it has been shown that for tasks that are easy, humans tend to make mistakes that are actually not caused by the modeling notation, but are rather the result of peculiarities of the human mind [10]. In order to overcome these problems and to improve validity of the collected data, we propose to additionally assess the subject's *mental effort*, i.e., the mental resources required for performing the task. We would like to stress that we *do not* propose replacing accuracy and duration, but rather using mental effort as an additional perspective that potentially provides further insights. The contribution of this paper is twofold. First, we argue for measuring mental effort on the basis of literature. Second, we will substantiate our claims with experiences drawn from own experiments.

*This research is supported by Austrian Science Fund (FWF): P23699-N23

The remainder of this paper is structured as follows. Section 2 discusses problems related to accuracy and how to address them using mental effort. Insights from experiments making use of mental effort are presented in Section 3 and afterwards discussed in Section 4. Section 5 presents related work and Section 6 concludes with a summary.

2. MEASURING MENTAL EFFORT

In the following we start by discussing the previously described problems in more detail. Then, we introduce *mental effort* to address the aforementioned problems.

Problems Concerning Accuracy. In the introduction, we briefly mentioned that accuracy is of rather low sensitivity and potentially incorrect for tasks of low complexity. Issues regarding the sensitivity become clear when looking at the definition of accuracy. Usually, accuracy is defined to be the ratio of correctly performed tasks (e.g., correct answers) divided by the number of all performed tasks (e.g., total amount of questions). In other words, subjects have to commit mistakes in order to obtain a lower accuracy. If a task performed in the course of an experiment is not difficult enough to provoke errors, no differences can be observed with respect to accuracy, also known as *ceiling effect* [25]. Likewise, if differences between experimental tasks are not large enough, no differences can be observed.

In addition, for tasks of low complexity a further problem arises—it has been recognized that for such tasks subjects tend to commit more careless mistakes. In [10], this phenomenon is explained by *Dual-Process Theory* [22]. Roughly speaking, this theory postulates that the human brain consists of two systems, *S1* and *S2*. S1 processes are characterized as being fast, unconscious and effortless. S2 processes, in contrast, are slow, conscious and effortful. In addition, S2 serves as monitor of the fast automatic responses of S1. Apparently, in certain situations, S1 comes up with a fast response and S2 does not intervene—leading to answers that are fast but incorrect. Hence, for tasks of low complexity, it may be the case that accuracy does not reflect the task's difficulty, but rather this peculiarity of the human mind.

Up to now we have discussed problems associated with measuring accuracy, i.e., low sensitivity and potential problems when assessing accuracy for tasks of low complexity. In the following, we introduce the concept of *mental effort* and illustrate how it can be used to overcome these problems.

Measuring Mental Effort. In general, the human brain can be seen as a *"truly generic problem solver"* [24]. Within the human brain, cognitive psychology differentiates between working memory that contains information currently being processed, as well as long-term memory in which information can be stored for a long period of time [17]. Most severe, and thus of high interest, are limitations of the working memory. As reported in [2], working memory cannot hold more than about seven items at the same time. The amount of working memory currently used is thereby referred to as *mental effort*. The importance of the working memory has been recognized and led to the development and establishment of Cognitive Load Theory, meanwhile widespread and empirically validated in numerous studies [3].

To measure mental effort, various techniques, such as rating scales, pupillary responses or heart-rate variability are available [17]. Especially rating scales, i.e., self-rating mental effort, has been shown to reliably measure mental effort and is thus widely adopted [9, 17]. Furthermore, this kind of measurement can easily be applied, e.g., by using 7-point rating scales. For instance, in [13] mental effort was assessed using a 7-point rating scale, ranging from *(1) very easy* to *(7) very hard* for the question *"How difficult was it for you to learn about lightning from the presentation you just saw?"*.

In the context of conceptual models, mental effort is of interest as it appears to be connected to performance, e.g., properly answering questions about a model. In general, it is known that errors are more likely to occur when the working memory's limits are exceeded [23]. Similarly, in [14] it is argued that higher mental effort is in general associated with lower understanding of models.

Based on these insights, we argue that mental effort is connected to performance, i.e., accuracy and duration. In contrast to accuracy, however, subtle differences can presumably be observed. In particular, for cases where mental effort is well within the working memory's limits and thus does not provoke a significant amount of errors, still a different mental effort could be observed. In addition, for tasks of low complexity, careless mistakes may distort the measurement of accuracy. For mental effort, however, it can be expected that careless mistakes will not distort the measurement.

3. MENTAL EFFORT IN EMPIRICAL RESEARCH

So far, our arguments for measuring mental effort are based on insights from literature. In the following, we will complement the discussion with findings we gained in own experiments. For each experiment, we will shortly sketch the setup on a very abstract level and point out relevant findings related to the measurement of mental effort.

3.1 Experiment 1: Test Cases in Declarative Business Process Modeling

Background. In this experiment, we investigated whether the presence of test cases supports the maintenance of declarative business process models, as argued in [32]. In the context of this work, the relevant information is that subjects were asked to adapt conceptual models with different types of operational support.

Setup. We employed a randomized, balanced single-factor experiment with repeated measurements (cf. [27]). The factor was *adoption of test cases*, having factor levels *test cases* as well as *absence of test cases*. The experiment's objects were change assignments for two declarative process models. Measured response variables relevant for this work were *mental effort* and *accuracy*, i.e., the amount of errors conducted (details are provided in [31]). To assess mental effort, we employed a 7-point rating scale, ranging from *Extremely low mental effort (1)* to *Extremely high mental effort (7)* for the question *"How would you assess the mental effort for completing the change tasks (with tests)?"*. For assessing the impact of factor level *absence of test cases*, the phrase *"with tests"* was replaced by *"without tests"*.

Execution of Experiment. The experiment was conducted in December 2010 in a course on business process management at the University of Innsbruck; in total 12 students participated. Operational support for collecting demographic data was provided by Cheetah Experimental Platform (CEP) [21], modeling assignments were conducted using Test Driven Modeling Suite (TDMS) [30].

Findings Relevant to Mental Effort. The collected data indicated that subjects, who had test cases at hand, conducted fewer errors, however, the differences were not significant (Wilcoxon Signed-Rank Test, Z = -0.857, p = 0.391). Interestingly, the data indicated a lower mental effort for subjects who had test cases at hand. However, in this case the differences could be found to be significant (Wilcoxon Signed-Rank Test, Z = -2.565, p = 0.010). A detailed analysis showed that the provided models were too simple to provoke the desired effects, i.e., differences with respect to the amount of errors committed. In fact, 96% of the tasks were performed correctly, leaving almost no room for improvement. Still, the models were difficult enough to achieve significant results with respect to mental effort. Knowing that errors are more likely to occur when the working memory's limits are exceeded [23], these results seem plausible. Even though the tasks were not difficult enough to go beyond the limit of the subjects' working memory and thereby provoking errors, different levels of mental effort were required. Put differently, it appears as if in this case measuring mental effort provided a more sensitive method than accuracy.

3.2 Experiment 2: Test Cases in Declarative Business Process Modeling (Replication)

Background. In this experiment, we further explored this research direction, i.e., the background is identical to Experiment 1.

Setup. Since this experiment is a replication of Experiment 1, the setup is identical, except for more complex models[1].

Execution of Experiment. The experiment was conducted in December 2011 in a course on business process management at the University of Ulm; in total 31 students participated. Again, CEP [21] and TDMS [30] were used as operational support.

Findings Relevant to Mental Effort. Data analysis showed that subjects who had test cases at hand conducted significantly less errors (Wilcoxon Signed-Rank Test, Z = -3.165, p = 0.002) and required significantly less mental effort (Wilcoxon Signed-Rank Test, Z = -3.867, p = 0.000). Interestingly, the total amount of correctly performed tasks dropped from 96% in Experiment 1 to 80% in this experiment. Hence, the two key observations are, as follows. First, apparently a certain level of complexity was required to provoke enough errors and to make differences with respect to accuracy significant. Second, mental effort consistently showed significant differences in both experiments. In other words, as discussed in Section 2, mental effort and accuracy seem connected, however, a certain level of complexity is required for accuracy in order to show significant differences.

[1]Material can be downloaded from:
http://bpm.q-e.at/experiment/TDMReplication

3.3 Experiment 3: Hierarchy in Business Process Models

Background. In this experiment we investigated the impact of hierarchy on the understandability of BPMN models. In the context of this work, the essential part is that we elaborated pairs of information-equivalent models, one of them making use of hierarchy. Then, we asked subjects to answer questions about those models, measuring accuracy of answers, duration and mental effort.

Setup. We employed a randomized, balanced single-factor experiment with repeated measurements (cf. [27]). The factor was *hierarchy* with factor levels *flat* as well as *hierarchical*. The experiment's objects were two BPMN-based business processes. Measured response variables relevant for this work were *accuracy of answers*, *duration* and *mental effort*[2]. In contrast to Experiment 1 and Experiment 2, where mental effort was assessed once for each subject, in this experiment we measured the expended mental effort after each question. To assess mental effort, we used a 7-point rating scale ranging from *Extremely low mental effort (1)* to *Extremely high mental effort (7)*. The question for measuring mental effort was: *"Please indicate your mental effort for answering this question (question X)"*.

Execution of Experiment. The experiment was conducted in a course on business process management at the University of Eindhoven in January 2012; in total 114 students participated. Again, CEP [21] was used for presenting the models to subjects and collecting answers.

Findings Relevant to Mental Effort. The assessment of accuracy, duration and mental effort *per question*, as opposed to Experiment 1 and Experiment 2, where mental effort was assessed once for the entire experiment, allowed for a much more fine grained analysis. In the course of this experiment, 2 BPMN-based business process models were presented to each subject. For each model, 8 questions were asked, leading a total of 16 questions per subject. Since we expected different mental effort, accuracy and duration, depending on whether a question was posed for a hierarchical model or a flat model, responses were analyzed separately, leading to a total of 32 data points. In the following, we will discuss this data from two perspectives. First, we will present a case in which accuracy did not reflect the difficulty of a task, but rather peculiarities of the human mind. Second, we will take a closer look into the relation between mental effort, accuracy and duration.

Accuracy for Tasks of Low Complexity. In Section 2 we argued that measurement of accuracy might lead to unexpected results—in the following, we provide further empirical evidence. In particular, the third question in this experiment yielded an average mental effort of 3.14, accuracy of 0.79 and duration of 40 seconds when asked for the hierarchical model. If the same question was posed for the information-equivalent model without hierarchy, mental effort increased to 3.75, duration increased to 51 seconds, but also the accuracy increased to 0.91. Statistically speaking, a Mann-Whitney U test showed that mental ef-

[2]Material can be downloaded from:
http://bpm.q-e.at/experiment/Hierarchy

fort increased significantly (z = -3.271, p = 0.001), also the duration increased significantly (z = -4.468, p = 0.000). Apparently inconsistently, also the average accuracy increased, even though not significantly (z = -1.717, p = 0.086)—according to previous findings, *higher* mental effort should have been associated with *lower* accuracy.

In order to resolve this contradiction, we investigated the aforementioned question in detail. The analysis showed that it should have been harder to answer the question for the non-hierarchical model, i.e., lower accuracy could be expected. In particular, for answering the question in the hierarchical model, 13 BPMN nodes had to be taken into account—for the non-hierarchical model, however, 92 nodes had to be considered[3]. Knowing that this amount of nodes required the subjects to scroll considerably to see all relevant model elements, it seems surprising that actually a *higher* accuracy could be observed. However, in the light of Dual-Process Theory [22], these results can be explained as follows. For the hierarchical model, the question could be answered easily, as indicated by the average mental effort of 3.14 (approximately *Low mental effort*). Hence, it seems plausible that system S1 provided a quick, but incorrect answer that was not overridden by S2. In the non-hierarchical model, subjects were forced to scroll to determine the answer, i.e., involving conscious activities, hence activating S2. The activation of S2, in turn, ensured that the question was answered in a controlled way, instead by a quick response of S1. Summarizing, it seems as if relying on accuracy would have been misleading in this case, while mental effort and duration provided more reliable results.

Validity of Mental Effort. So far we have provided empirical evidence that mental effort is more sensitive than accuracy and can be measured properly for tasks of low complexity. In the following, we will provide empirical evidence that mental effort is tightly connected to accuracy and duration, i.e., is a valid measure of performance. To visualize the interplay between mental effort and accuracy, we used a scatter plot (cf. Figure 1). Therein, the x-axis represents the average mental effort required for answering a question. Values from 1 to 7 represent the rating scale used for assessing mental effort, ranging from *Extremely low mental effort (1)* to *Extremely high mental effort (7)*. Likewise, the y-axis reflects a question's average accuracy, i.e., the ratio of correct answers to total answers given for a question. As discussed in Section 2, higher mental effort should be associated with lower performance. Hence, in this particular case, higher mental effort should be associated with lower accuracy. In fact, in Figure 1, a tendency toward lower accuracy with increased mental effort can be observed. In particular, the dashed line represents the regression line as obtained through simple linear regression (R^2 = 0.41, F(1,30) = 21.16, p = 0.000). Please note that this regression does not contradict the case when mental effort and accuracy do not perfectly correlate, as demonstrated in the example above. Rather, the regression is not perfect, hence leaving room for such idiosyncrasies.

Akin to Figure 1, in Figure 2, the interplay between mental effort and duration is illustrated. Likewise, the x-axis

[3]The models and question can be accessed through: http://bpm.q-e.at/misc/HierarchyQuestion3

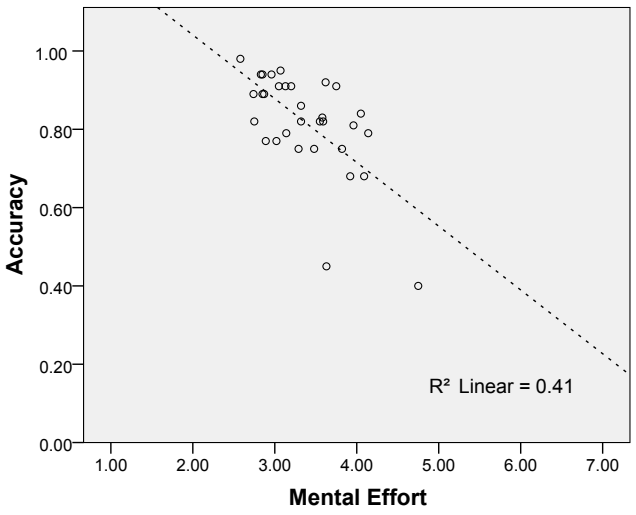

Figure 1: Mental effort versus accuracy

represents mental effort. On the y-axis, the average amount of seconds required for answering a question can be found. The dashed line is the result of simple linear regression (R^2 = 0.55, F(1,30) = 36.70, p = 0.000). Interestingly, in this case higher mental effort is associated with higher duration. In the light of the background presented in Section 2, also this result seems plausible. The more difficult a questions is to answer, the longer the answering process will take and the higher the mental effort will be.

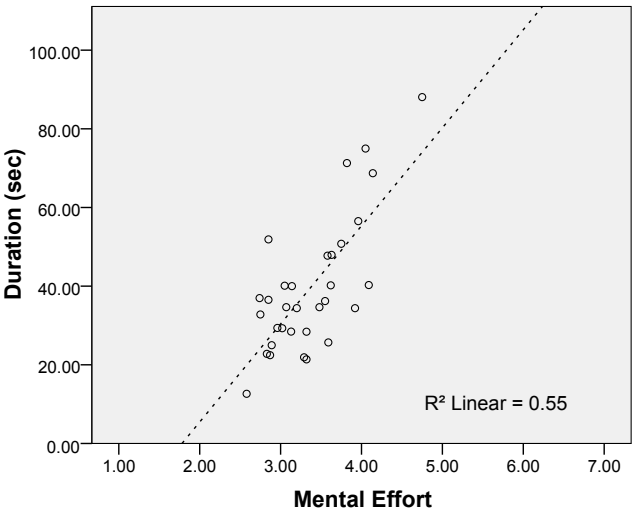

Figure 2: Mental effort versus duration (sec)

To substantiate these observations, we computed Pearson Correlation coefficient for correlations between mental effort, accuracy and duration. As shown in Table 1, the findings coincide with the observations made in Figures 1 and 2. In particular, the results confirm that mental effort and accuracy are correlated negatively (r(30) = -0.643, p = 0.000); mental effort and duration are correlated positively (r(30) = 0.742, p = 0.000). Finally, accuracy and duration are

correlated negatively (r(30) = -0.459, p = 0.008).

		Mental Eff.	Duration
Accuracy	Pearson Corr.	-0.643**	-0.459**
	Sig. (2-tailed)	0.000	0.008
	N	32	32
Mental Eff.	Pearson Corr.		0.742**
	Sig. (2-tailed)		0.000
	N		32

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

Table 1: Correlations

4. DISCUSSION

Up to know we argued that accuracy is presumably rather insensitive and may be distorted for tasks of low complexity. In order to tackle these problems, the measurement of mental effort was proposed. In the following, key insights as well as limitations of this approach are discussed.

Regarding sensitivity, Experiment 1 and Experiment 2 provided empirical evidence that mental effort is more sensitive than accuracy. In Experiment 1 tasks of rather low complexity were provided to the subjects. Even though differences with respect to accuracy and mental effort could be observed, only mental effort differed significantly [31]. In Experiment 2 the task complexity was increased, consequently more errors were committed. Knowing that errors are more likely to be committed when the working memory is overloaded [23], these observations seem plausible. In Experiment 1, different levels of mental effort could be observed. However, the working memory was not overloaded, resulting in a low error rate and hence marginally fluctuations in accuracy. In Experiment 2, increased complexity lead to an overload of working memory, accordingly the error rate increased and accuracy dropped. In other words, it seems likely that differences with respect to mental effort can be observed before differences with respect to accuracy occur, i.e., mental effort appears to be more sensitive.

Regarding tasks of low complexity, Experiment 3 provided further insights. In particular, we could observe an increase of mental effort and duration that was connected to increased accuracy—actually a decrease of accuracy could be expected, as far more model elements had to be taken into account. As indicated in [10], it seems as if this result can be traced back to peculiarities of the human mind, which tends to commit more careless mistakes for tasks of low complexity. Hence, in such a case it seems as if the measurement of mental effort provides a more accurate picture. Please note that this finding does not contradict the correlation between mental effort and accuracy, as found in Experiment 3. Rather, the correlation is valid in general, while this peculiar interplay could be found for one specific question.

Apparently, several limitations apply to this work. First, as shown in Figure 2, a *linear relationship* between mental effort and duration could be found. Even though this seems plausible for short-lasting tasks (the maximum average duration was about 90 seconds), it seems questionable in how far this holds for longer tasks. For instance, a long-lasting repetitive task, such as finding all elements named *"A"* within a model, will most likely lead to a low mental effort, *but* a long duration. Second, mental effort is a subjective measure. Even though it has been shown that people are able to give a numerical indication of their mental burden [16], it is questionable whether mental effort of different subjects can be compared directly. Hence, it seems advisable to ensure a reasonable sample size when conducting between-subject experiments or to focus on within-subject experiments. Third, we reported from consistent results across three experiments. Still, our findings may be peculiarities of these experiments. To improve the generalization, more experiments adopting different modeling languages are required.

5. RELATED WORK

In the domain of cognitive psychology, the work of Paas et al., in which mental effort is discussed broadly, is directly connected [17]. In contrast to this work, however, mental effort is not linked to conceptual modeling. In the domain of conceptual modeling, related work can be found where model understandability is of concern. For instance, Aranda et al. provide guidelines for assessing model understandability [1]. Besides accuracy and duration, *perceived difficulty* is proposed to be measured. However, in contrast to this work, no indications on how perceived difficulty can be measured, are given. Likewise, [11] assesses in a survey how understandability of models is measured. The most prominent measure is accuracy, followed by duration and perceived ease of understanding. However, these studies rather rely on the ease-of-use scale from *Technology Acceptance Model* [6] than on rating scales for measuring mental effort. Recently, mental effort has also been used as a tool for discussing model understandability. For instance, in [29] the role of mental effort in Business Process Modeling is discussed. Likewise, in [28, 33] mental effort is employed for assessing the impact of hierarchy on model understandability. In contrast to this work, however, mental effort is rather used as a tool for discussion; the measurement of mental effort is not of concern. Apparently, measuring mental effort is only meaningful if the validity of the experimental design can be ensured. In this respect [8, 18] provide guidelines for the proper operationalization of measurements.

6. SUMMARY AND OUTLOOK

In this work, we showed how measuring mental effort allows to compensate for shortcomings when measuring accuracy. In particular, we argued that mental effort is more sensitive than accuracy and that the measurement is not distorted for tasks of low complexity. Hence, it allows to identify subtle differences between conceptual models or conceptual modeling languages. Likewise, when data regarding accuracy is affected by ceiling effects, mental effort can still provide valuable insights. In addition, we showed that the measurement of mental effort can be implemented easily through the adoption of rating scales. Thereby, we recommend to measure mental effort after *each* task in order to provide a fine-grained measurement. With this contribution we hope to help in paving avenues for even more comprehensive empirical investigations.

Future work will imply the collection of further data for a deeper investigation of the interplay between mental effort, accuracy and duration. In particular, we plan to adopt eye

movement analysis [19] to constantly monitor mental effort while subjects perform a task.

7. REFERENCES

[1] J. Aranda, N. Ernst, J. Horkoff, and S. Easterbrook. A Framework for Empirical Evaluation of Model Comprehensibility. In *Proc. MISE '07*, pages 7–12, 2007.

[2] A. Baddeley. Working Memory: Theories, Models, and Controversies. *Annu. Rev. Psychol.*, 63(1):1–29, 2012.

[3] M. Bannert. Managing cognitive load—recent trends in cognitive load theory. *Learning and Instruction*, 12(1):139–146, 2002.

[4] J. C. Carver, E. Syriani, and J. Gray. Assessing the frequency of empirical evaluation in software modeling research. In *Proc. EESSMod '11*, pages 28–37, 2011.

[5] J. A. Cruz-Lemus, M. Genero, M. E. Manso, S. Morasca, and M. Piattini. Assessing the understandability of UML statechart diagrams with composite states—A family of empirical studies. *Empirical Software Engineering*, 25(6):685–719, 2009.

[6] F. Davies. *A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results*. PhD thesis, Sloan School of Management, 1986.

[7] A. M. Férnandez-Sáez, M. Genero, and M. R. V. Chaudron. Does the level of detail of uml models affect the maintainability of source code? In *Proc. EESSMod '11*, pages 3–17, 2011.

[8] A. Gemino and Y. Wand. A framework for empirical evaluation of conceptual modeling techniques. *Requir. Eng.*, 9(4):248–260, 2004.

[9] D. Gopher and R. Brown. On the psychophysics of workload: Why bother with subjective measure? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 26(5):519–532, 1984.

[10] I. Hadar and U. Leron. How intuitive is object-oriented design? *Communications of the ACM*, 51(5):41–46, 2008.

[11] C. Houy, P. Fettke, and P. Loos. Understanding understandability of conceptual models: What are we actually talking about? In *Proc. ER '12*, pages 64–77, 2012.

[12] R. Laue and A. Gadatsch. Measuring the Understandability of Business Process Models - Are We Asking the Right Questions? In *Proc. BPD '10*, pages 37–48, 2011.

[13] R. Mayer and P. Chandler. When learning is just a click away: Does simple user interaction foster deeper understanding of multimedia messages. *Journal of Educational Psychology*, 93(2):390–397, 2001.

[14] D. L. Moody. Cognitive Load Effects on End User Understanding of Conceptual Models: An Experimental Analysis. In *Proc. ADBIS '04*, pages 129–143, 2004.

[15] J. Mylopoulos. Information modeling in the time of the revolution. *Information Systems*, 23(3/4):127–155, 1998.

[16] F. Paas. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, 84(4):429–434, 1992.

[17] F. Paas, A. Renkl, and J. Sweller. Cognitive Load Theory and Instructional Design: Recent Developments. *Educational Psychologist*, 38(1):1–4, 2003.

[18] J. Parsons and L. Cole. What do the pictures mean? Guidelines for experimental evaluation of representation fidelity in diagrammatical conceptual modeling techniques. *DKE*, 55(3):327–342, 2005.

[19] J. Pinggera, M. Furtner, M. Martini, P. Sachse, K. Reiter, S. Zugal, and B. Weber. Investigating the Process of Process Modeling with Eye Movement Analysis. In *Proc. ER-BPM '12*, to appear.

[20] J. Pinggera, P. Soffer, S. Zugal, B. Weber, M. Weidlich, D. Fahland, H. Reijers, and J. Mendling. Modeling Styles in Business Process Modeling. In *Proc. BPMDS '12*, pages 151–166, 2012.

[21] J. Pinggera, S. Zugal, and B. Weber. Investigating the process of process modeling with cheetah experimental platform. In *Proc. ER-POIS '10*, pages 13–18, 2010.

[22] K. E. Stanovich and R. West. Individual differences in reasoning: implications for the rationality debate? *Behavioural and Brain Sciences*, 23(5):665–726, 2000.

[23] J. Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988.

[24] W. J. Tracz. Computer programming and the human thought process. *Software: Practice and Experience*, 9(2):127–137, 1979.

[25] W. P. Vogt. *Dictionary of Statistics & Methodology: A Nontechnical Guide for the Social Sciences (Fourth Edition)*. SAGE Publications, 2011.

[26] C. Wohlin, M. Höst, and K. Henningsson. Empirical research methods in software engineering. In *Empirical Methods and Studies in Software Engineering*, volume 2765 of *LNCS*, pages 7–23. Springer, 2003.

[27] C. Wohlin, R. Runeson, M. Halst, M. Ohlsson, B. Regnell, and A. Wesslen. *Experimentation in Software Engineering: an Introduction*. Kluwer, 2000.

[28] S. Zugal, J. Pinggera, J. Mendling, H. Reijers, and B. Weber. Assessing the Impact of Hierarchy on Model Understandability—A Cognitive Perspective. In *Proc. EESSMod '11*, pages 123–133, 2011.

[29] S. Zugal, J. Pinggera, and B. Weber. Assessing process models with cognitive psychology. In *Proc. EMISA '11*, pages 177–182, 2011.

[30] S. Zugal, J. Pinggera, and B. Weber. Creating Declarative Process Models Using Test Driven Modeling Suite. In *Proc. CAiSE Forum '11*, pages 16–32, 2011.

[31] S. Zugal, J. Pinggera, and B. Weber. The impact of testcases on the maintainability of declarative process models. In *Proc. BPMDS '11*, pages 163–177, 2011.

[32] S. Zugal, J. Pinggera, and B. Weber. Toward Enhanced Life-Cycle Support for Declarative Processes. *Journal of Software: Evolution and Process*, 24(3):285–302, 2012.

[33] S. Zugal, P. Soffer, J. Pinggera, and B. Weber. Expressiveness and Understandability Considerations of Hierarchy in Declarative Business Process Models. In *Proc. BPMDS '12*, pages 167–181, 2012.