The effect of tangible media on individuals in business process modeling - a controlled experiment

Alexander Lübbe

Technische Berichte Nr. 41

des Hasso-Plattner-Instituts für
Softwaresystemtechnik
an der Universität Potsdam

Technische Berichte des Hasso-Plattner-Instituts für
Softwaresystemtechnik an der Universität Potsdam

Technische Berichte des Hasso-Plattner-Instituts für Softwaresystemtechnik an der Universität Potsdam | 41

Alexander Lübbe

The effect of tangible media on individuals in business process modeling

A controlled experiment

Universitätsverlag Potsdam

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Universitätsverlag Potsdam 2011
http://info.ub.uni-potsdam.de/verlag.htm

Am Neuen Palais 10, 14469 Potsdam
Tel.: +49 (0)331 977 4623 / Fax: 3474
E-Mail: verlag@uni-potsdam.de

Die Schriftenreihe **Technische Berichte des Hasso-Plattner-Instituts für Softwaresystemtechnik an der Universität Potsdam** wird herausgegeben von den Professoren des Hasso-Plattner-Instituts für Softwaresystemtechnik an der Universität Potsdam.

ISSN (print) 1613-5652
ISSN (online) 2191-1665

Online veröffentlicht auf dem Publikationsserver der Universität Potsdam
URL http://pub.ub.uni-potsdam.de/volltexte/2011/4900/
URN urn:nbn:de:kobv:517-opus-49001
http://nbn-resolving.org/urn:nbn:de:kobv:517-opus-49001

Zugleich gedruckt erschienen im Universitätsverlag Potsdam:
ISBN 978-3-86956-108-0

The effect of tangible media on individuals in business process modeling – a controlled experiment

Alexander Luebbe and Mathias Weske

Hasso Plattner Institute, University of Potsdam, Germany
{alexander.luebbe,mathias.weske}@hpi.uni-potsdam.de
http://bpt.hpi.uni-potsdam.de

Abstract. In current practice, business processes modeling is done by trained method experts. Domain experts are interviewed to elicit their process information but not involved in modeling. We created a haptic toolkit for process modeling that can be used in process elicitation sessions with domain experts. We hypothesize that this leads to more effective process elicitation.
This paper brakes down "effective elicitation" to 14 operationalized hypotheses. They are assessed in a controlled experiment using questionnaires, process model feedback tests and video analysis. The experiment compares our approach to structured interviews in a repeated measurement design. We executed the experiment with 17 student clerks from a trade school. They represent potential users of the tool. Six out of fourteen hypotheses showed significant difference due to the method applied. Subjects reported more fun and more insights into process modeling with tangible media. Video analysis showed significantly more reviews and corrections applied during process elicitation. Moreover, people take more time to talk and think about their processes.
We conclude that tangible media creates a different working mode for people in process elicitation with fun, new insights and instant feedback on preliminary results.

Key words: Process Modeling, tangible media, individuals, process elicitation, BPMN, t.BPM, controlled experiment

1 Introduction

Business process management can be seen as a management approach to structure work in organizations [1]. In the last decade the term was coined as an IT approach to support or automate working procedures in organizations using software systems [2]. Supporting processes with software offers great potential to save time, enhance reliability and deliver standardized output [3, 4]. However, implementing a process in a heterogenous software environment requires significant software engineering effort. As in all software projects, misunderstandings in early stages lead to expensive change requests at later stages of projects [5]. Thus, the quality

of communication between stakeholders is crucial to translate user and system requirements into software implementation [6].

In business process management graphically depicted process models serve as communication vehicle about working procedures. These models are created by specially trained method experts, typically external consultants. They gather the knowledge for these models in interviews or workshops [7, 8] with the stakeholders of the process. Afterwards, the method expert creates a business process model using notations such as EPC [9] or BPMN [10]. Modeling is a process of filtering and framing the information gathered. Drawing of process models is supported with specialized software which also implies specialized knowledge to use it. The domain expert, a stakeholder of the process, is then asked to provide feedback to a printed process model.

Domain experts, in many cases are unfamiliar with process-orientation. Because they often don't have sufficient exposure and training to understand process modeling concepts, they are unable to fully understand the models. If they conclude that their knowledge is not appropriately represented, additional effort is needed to explain the model and to resolve misunderstandings.

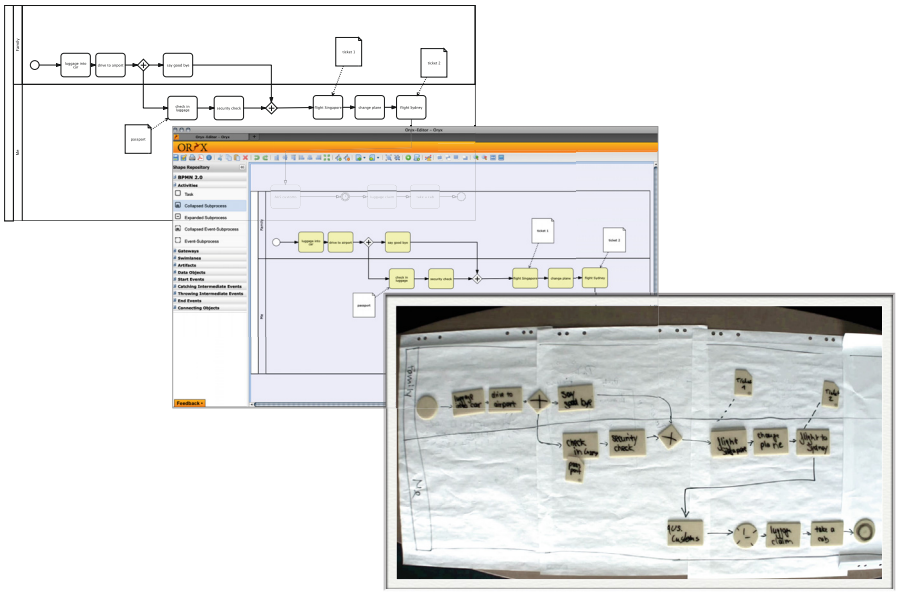


Fig. 1. The same process: as a printed model (upper left), in a software modeling tool (middle) and modeled with t.BPM (bottom right)

We have developed the tangible Business Process Modeling (t.BPM) Toolkit, see Figure 1. It is a transcribable set of plastic tiles that can be used to model processes on a table. It reflects the iconography of the Business Process Modeling

Notation [10] (BPMN). In our opinion, it enables domain experts to model their processes and allows the method expert to act as a facilitator. For the scope of our work, we consider domain experts to be the stakeholders of the project, i.e. clerks or managers. The method expert is either an external process consultant or an internal process expert who is trained in methods and notations to frame knowledge in process-oriented projects.

We anticipate t.BPM to be used in group modeling sessions instead of post-its, brown paper or software tools. Moreover, we think that one-to-one interview situations gain from complementing them with t.BPM. This paper, reports on a controlled experiment in which we analyze the effect of t.BPM in one-to-one interview situations with respect of the effectiveness of process elicitation with or without t.BPM.

In Section 2 we investigate related work. In Section 3 we explain the hypotheses, the experiment setup, the variables and the analysis procedures used in this work. The experiment execution is explained in Section 4. The data analyzation is reported in Section 5. The results from the analysis are discussed in Section 6. Finally, we conclude the paper in Section 7.

2 Related Work

Empirical research on process modeling is typically focussed on the models that were produced with software tools and can be automatically analyzed [11, 12, 13]. Only recently, BPM research also turned towards the modelers [14, 15, 16] in front of the screen and the process of model creation [17, 18, 19].

To give some examples, Recker investigated the relation between modeling grammars and modeler performance. He found that modeler performance is influenced by the complexity of the grammar at use [17] as well as personal factors such as modeling experience and modeling background [20]. Sedera et. al. [18] used case study research and survey methods to derive qualitatively a framework of factors that influence the success of process modeling efforts in companies such as user participation, modeling methodology and tooling.

Rittgen sees process modeling as a negotiation [21, 22, 23]. Using a combination of design science [24] and action research [25], he developed a software tool and a method to guide groups in workshops [19, 26]. The method contains six steps and is strongly supported by the software that was built for it.

Controlled experiments for process modeling have been conducted by e.g. Weidlich [27], Weber [28] and Holschke [15] to investigate the influence of change request types on model quality[27], the effect of events on planing performance [28], or model granularity on reusability of artifacts [15]. To our best knowledge there is no controlled experiment that investigates the presence of a mapping tool for business process modeling.

Tangibility as a quality for interaction is studied in multiple disciplines such as HCI [29] or industrial design [30]. In design research, which is the scientific investigation of the design process through cognitive, qualitative or ethnographic

methods [31], tangible prototyping is seen as a key enabler to collect feedback to ideas in early design stages [32, 33]. The embodiment of an idea as an intermediate representation permits distributed cognition and either allows or prohibits access to collaborative creation [34]. In that context, Gibson coined the term affordances [35] as the attributes that indicate possible actions.

With t.BPM we transport the affordances of tangibility to the design process of process modeling. t.BPM affords actions which people can naturally identify and execute, such as pointing, moving or taking away. Collaborative creation is therefore not limited by access to the intermediate representation that is shared amongst the designers of the process.

Process modeling practice in the field has developed various ideas for model building in conjunction with end users. This typically happens in moderated groups [7, 36, 37] in which a modeling expert translates the input into a model that is discussed with the audience. A popular low tech version is brown paper modeling [4, 19, 38] in workshops in which the process is taped to the wall using post-its and differently shaped paper. However, there is no commonly agreed operationalization of this idea and the expected effects of this method. Unfortunately, process elicitation techniques practiced in the field are barely published as they are intellectual property of the consulting company that runs it. One exception is Unity[1]. The company uses the proprietary OMEGA process modeling method [39, 40] embedded in a "strategic production management" approach. Their best-practice suggests to use paper cards that reflect the iconography of the modeling elements in workshops or interviews. Cards are available in different sizes and the use is said to be depending on the consultants "gusto" and "experience" [41]. This method is said to have in general a "stimulating effect" on the participants, however a more comprehensible investigation is not presented.

In summary, some process modeling approaches point into similar directions, such as instant mapping or strong user involvement. Details about the facilitation or the effect are barely published. To our best knowledge, nobody has scientifically investigated the effect of a process mapping tool on the individual domain expert. Design research suggests that tangibility as an affordance lowers barriers for participation. In the case of t.BPM, this is of particular interest as we assume the domain experts to be the modelers driving the process creation. The setup and execution of our controlled experiment was guided by Creswell [42] and Wohlin [43]. We use literature from software engineering [44], psychology [45] and statistics [46] to inform the structure of the paper and the level of reporting.

[1] http://www.unity.de/

3 Experiment Planning

3.1 Goal and Hypotheses

Our goal is to examine the effect of t.BPM compared to structured interviews with single individuals. Structured interviews are seen as the most effective requirements elicitation technique [8]. By ?effective? we mean that it produces a ?desired or intended result? [47]. In requirements engineering, more information is seem as more effective elicitation. It was already shown that visual representations not necessarily create more information [8]. We think effective process elicitation has more dimensions such as user engagement, iterated (higher quality) results and better feedback on process models. We decompose these areas further into fourteen hypotheses which we operationalize in Section 3.5. An overview of our hypotheses tree is given in Figure 3. The following considerations led to this hypothesis decomposition.

More user engagement. For decades HCI research investigates tangible interfaces [29] and as one factor the impact on task engagement. In those cases, engagement is typically measured as time spent on a problem [48]. We therefore also measure time and hypothesize that people will *spent more time talking* about the process but also *spent more time to think* about what they do. We do not hypothesize about the overall time because we assume that t.BPM will consume additional time to handle the shapes in comparison to interviews. Therefore we hypothesize about the time slices that we are most interested in.

Schaufeli developed different instruments to measure work engagement which he sees as the opposite of a burnout [49]. For him, work engagement has two dimensions, activation and identification [50]. One can argue that activation is already measured with the time spent on the task. We additionally hypothesize that people have *more fun* and have *more motivation* to accomplish the task which is another aspect of activation. The aspect of identification inspires us to hypothesize that people modeling with t.BPM are *more committed to the solution* that they shaped. That also means, they would have a *clearer goal* understanding of what they are doing, which we hypothesize.

Better information from elicitation. The cognitive load theory [51] postulates that our brain has limited capacity, called work memory. The fundamental insight was first reported by Miller in 1956 who found that people can hold on to "seven, plus or minus two" [52] information pieces at a time without context. The amount of information to be kept in the work memory can be reduced by externalizing knowledge [53] as it is done with t.BPM or other mapping approaches. Reduced load on working memory enables people to get into details more extensively. Thus, we hypothesize that people share more detailed process knowledge such as *more problems* with and *more phases* in the process when using t.BPM.

But better information does not simply mean more information. The quality of the initial workshop can be measured by the amount of iterations needed to agree on the result afterwards. Typically, consultants elicit a process in the workshop,

model it afterwards, and then send it out for people to review it, approve it or propose corrections [54]. In t.BPM the result of the initial elicitation workshop can be seen as a reviewed result. Since information is immediately mapped and framed, it provokes instant feedback [55]. We hypothesize that people will do *more reviews* of the process model and apply *more corrections* to their initially elicited story when using t.BPM due to the mapping effect.

Better feedback on process models. We strongly believe that better feedback is grounded in a deeper understanding. It is suggested that students who actively engage with the material are more likely to recall information afterwards [56]. Recall is the first stage of understanding before retention and generation [57]. Consequently, we hypothesize that people doing t.BPM have *new insights into process thinking* due to their hands-on experience. Better understanding should also enable people to read and understand models better.

Understandability tests for process models are at their early stage [58, 59]. Thus, we hypothesize about the positive effect to be expected in the field, e.g. that people with t.BPM experience will *find more mistakes* and *provide more comments* to process models when asked for feedback. Furthermore, we do think that better understanding will lead to *more commitment to feedback* and therefore hypothesize this as an indicator for the understanding that people build.

3.2 Experiment Setup & Sampling Strategy

We design the following experimental setup, see also Figure 2. Subjects get first conditioned to a certain level of BPM understanding. Therefore, we use a two page introduction and a sample model that explains how to make pasta. After conditioning, subjects are randomly assigned to do either interviews or model with t.BPM. The topic is randomly chosen between buying expensive equipment and running a call for tender. A structured questionnaire guides the experimenter through the experimental task. Two experimenters operate the experiment. One guides the subjects in the role of an interviewer, the other experimenter observes the situation and ensures a stable treatment throughout the experiment. They randomly swap roles.

During the experimental task data is collected using video recording. Afterwards, an eighteen item questionnaire is to be filled in by the subjects. Moreover, a sheet with a process model is handed to subjects. They are asked to provide feedback to process models that depict "finding a new flat" or "getting a new job", chosen randomly. In every step of the experiment, the time is tracked but time constrains are not imposed on subjects. After the first run, subjects rerun the experimental task using the other method and the other process to report on. They do the questionnaire the second time and get the other process model to provide feedback to.

In other words, the sampling strategy is a randomized balanced single factor design with repeated measurements [43] also known as a within-subjects design [60]. All subjects get both treatments assigned in different order. All subjects do interviews and process modeling. And all subjects get both processes to report

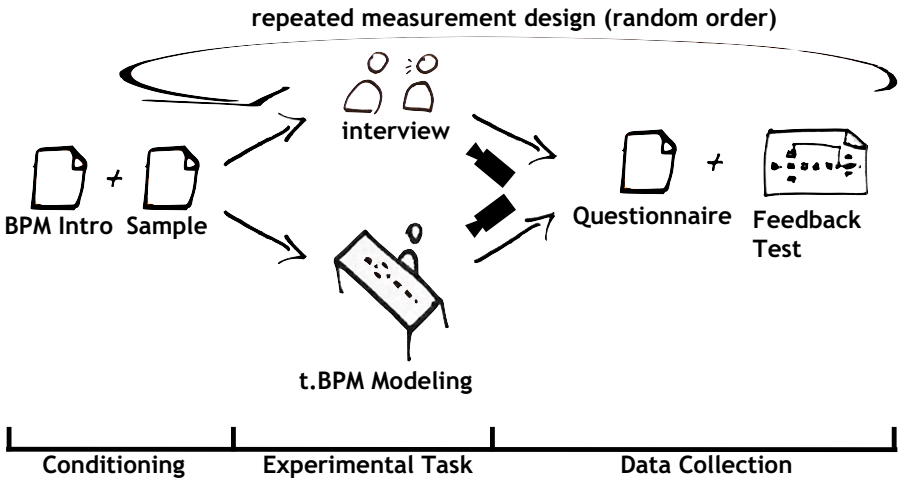


Fig. 2. Experiment Setup for this study

on and both feedback tests, again randomly assigned. Subjects are rewarded for their participation with a chocolate bar and a cinema voucher.

3.3 Experimental Material

We outline and explain the printed experimental material here. The original documents are appended to this paper, see Appendix A. Like the experiment, the experimental material is in german.

- BPM Introduction
  A two page document explaining the terms Business Process Management, Business Process Modeling and process models. The document can be found in the Figure 7 and Figure 8 in Appendix A.
- Sample Model
  A one page document that depicts the process of "Making Pasta". It also contains a legend of the BPMN elements used and four pragmatical hints on process modeling. In particular, it suggests the balanced use of gateways, an eighty percent rule for relevance to set granularity, verb-object style activity labels as suggested by Mendling et al. [16] and a notational convention for conditions at gateways. The document can be found in Figure 9 in Appendix A.
- Task Sheet
  One paragraph explaining the experimental task. Subjects are asked to model or report on one of the following processes: buying a new flat screen for the entrance to the company building or running a call for tenders to build a new warehouse. The introduction explicitly sets the context, the start and the end-point of the process. The task sheets can be found in Figure 10 in Appendix A.

- Interview Guide (for Experimenter)
  Experimenters guided through the modeling / interview by asking the same six questions in the same order in the experimental task. It started with "Please identify all relevant steps", went on with "Which documents play a role?" and concluded with "Which problems are you expecting in this process" and "Is there anything else you want to tell us about the process?". Experimenters read out the exact questions from the interview guide. It also contains standardized answers to questions from participants, such as "Make an assumption and proceed from there". The interview guide can be found in Figure 11 in Appendix A.
- Questionnaire
  Closed questionnaire with eighteen questions to be rated on a 5-point Likert scale. Three questions operationalize one hypothesis. For more information on this see Section 3.5. The questionnaire can be found in Figure 12 in Appendix A.
- Feedback Test
  A process model, a sample annotation and the request to "provide feedback" to the model. Two versions of this test exist. One on "Moving to a new flat" and another one on "Getting a new job". The process models contained problems which we intentionally build into them. More details on this can be found in [54]. The feedback tests can be found in Figure 13 and Figure 14 in Appendix A.

3.4 Participant Selection

The sample population, used in research studies, should be representatives of the population to which the researchers wish to generalize [61]. Thus, we want potential users of t.BPM to participate in the study. Clerks were identified as the most suitable group. They run processes on an operational level and might be questioned in business process elicitation projects as stakeholders of the processes.

We contacted a trade school in Potsdam (Germany) and got access to run the experiment on-site. Amongst other professions the trade school educates office and industrial clerks. Industrial clerks do planing, execution and controlling of business activities. Office clerks do supporting activities in a department, e.g. as office managers. On the job, both professions might overlap depending on the size of the company. We decided that the subjects very well represent the target population.

3.5 Operationalized Hypotheses

Using the video material, the questionnaires and the feedback test we operationalize the hypotheses presented in Section 3.1. Figure 3 provides an overview about the operationalized hypotheses. In the following sections we define each hypothesis as H_{1xx} and its null hypothesis as H_{0xx}.

Questionnaire Hypotheses (H_{111},H_{112},H_{113},H_{116}, H_{131}, H_{132}) Hypotheses which rely on perceived measures are tested using a questionnaire.

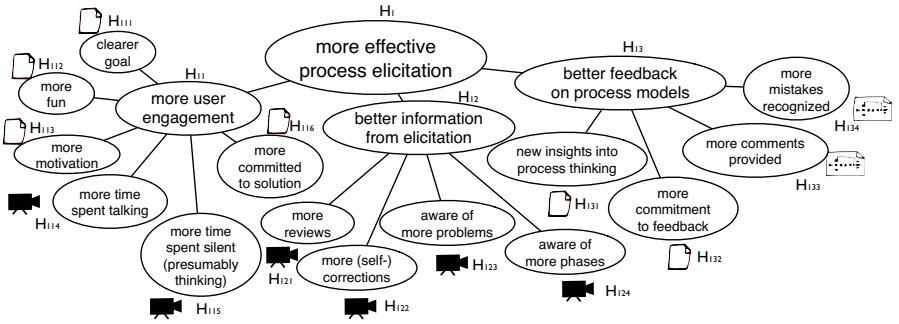


Fig. 3. Overview of the operationalized Hypotheses

On a five-point Likert scale subjects rate their agreement to, in summary, eighteen statements. Each hypothesis is tested by presenting three statements. Two statements are formulated towards the hypotheses, one is negatively formulated. The level of agreement is mapped to the values one to five where one is no agreement and five is a strong agreement. The values are aggregated (negative statement is turned around by calculating $6 - value$) to retrieve the actual value to work with. The hypothesis holds if there is a significant difference according to the method immediately used before, t.BPM or interviews. Accordingly, we define

- $Q = (q_1, .., q_{18})$, i.e. the sequence of statements in the questionnaire
- $p : Q \rightarrow [1, 2, 3, 4, 5]$, i.e. the mapping function that assigns a value the statements from the questionnaire
- $P = p_1, ..., p_n$, i.e. the set of all mapping functions (one per filled-in questionnaire)
- $P = P_{tbpm} \bigcup P_{int}; P_{tbpm} \bigcap P_{int} = \emptyset$
- $h := \mathcal{P}(P) \times Q \times Q \times Q = \mathcal{P}(P) \times Q^3$
- $h(P, x, y, z) = \frac{\sum_{\forall p \in P} p(x) + p(y) + (6 - p(z))}{3 * |P|}$

To calculate the average per hypothesis we define function h, see above. As input, it takes a set of mappings (filled-in questionnaires) and three questions that should be aggregated to represent the value for one hypothesis. By convention the last variable z is always the negatively wired statement from the questionnaire. By using function h with the P_{tbpm} and P_{int}, the disjunct sets of mapping functions for t.BPM and interview sessions, we can define the Hypotheses in the following way:

- H_{111}:$h(P_{tbpm}, 6, 15, 8) > h(P_{int}, 6, 15, 8)$, i.e. subjects report a clearer goal understanding in t.BPM sessions than in interviews.
  H_{011}:$h(P_{tbpm}, 6, 15, 8) \leq h(P_{int}, 6, 15, 8)$, , i.e. subjects report equal or less clarity in goal understanding for t.BPM sessions.

- H_{112}:$h(P_{tbpm}, 2, 14, 18) > h(P_{int}, 2, 14, 18)$, i.e. subjects report more fun in t.BPM sessions than in interviews.
  H_{012}:$h(P_{tbpm}, 2, 14, 18) \leq h(P_{int}, 2, 14, 18)$, i.e. subjects report equal or less fun in t.BPM sessions.
- H_{113}: $h(P_{tbpm}, 4, 11, 7) > h(P_{int}, 4, 11, 7)$, i.e. subjects report to be more motivated in t.BPM sessions than in interviews.
  H_{013}: $h(P_{tbpm}, 4, 11, 7) \leq h(P_{int}, 4, 11, 7)$, i.e. subjects report to be more equally or less motivated in t.BPM sessions.
- H_{116}: $h(P_{tbpm}, 15, 17, 5) > h(P_{int}, 15, 17, 5)$, i.e. subjects report to be more committed to the solution of t.BPM sessions than in interviews.
  H_{016}: $h(P_{tbpm}, 15, 17, 5) \leq h(P_{int}, 15, 17, 5)$, i.e. subjects report to be equally or less committed to the solution of t.BPM sessions.
- H_{131}: $h(P_{tbpm}, 9, 12, 3) > h(P_{int}, 9, 12, 3)$, i.e. subjects report to gain more new insights to process understanding from t.BPM sessions than from interviews.
  H_{031}: $h(P_{tbpm}, 9, 12, 3) \leq h(P_{int}, 9, 12, 3)$, i.e. subjects report to gain equally or less new insights to process understanding from t.BPM sessions.
- H_{132}: $h(P_{tbpm}, 1, 10, 16) > h(P_{int}, 1, 10, 16)$, i.e. subjects report to be more committed to feedback on process models after t.BPM sessions than after interviews.
  H_{032}: $h(P_{tbpm}, 1, 10, 16) \leq h(P_{int}, 1, 10, 16)$, i.e. subjects report to be equally or less committed to feedback on process models after t.BPM sessions.

For details, you find the document in Figure 12 in Appendix A.

Video Hypotheses (H_{114},H_{115},H_{121},H_{122}, H_{123}, H_{124}) We operationalize hypotheses related to time and actions taken during the experimental task using video coding analysis. Therefore we define the following coding schemes:

- **Time Slicing(H_{114},H_{115})**: The duration of the experimental task is sliced exclusively to belong to one of the five categories. The Use_{tBPM} of t.BPM such as labeling and positioning the shapes without talking, $Talk_{tBPM/int}$ is the time people talk about the process, $UseTalk_{tBPM}$ is talking and using t.BPM (to avoid overlap between Use_{tBPM} and $Talk_{tBPM}$), $Silence_{tBPM/int}$ is the time spent silent, e.g. when people do not talk and do not handle t.BPM. Finally, $Rest_{tBPM/int}$ captures remaining time such as interactions with the interviewer. The same coding scheme is used for both experimental tasks. However, Use and $UseTalk$ do not apply for interviews as there is no t.BPM to use.
- **Corrections and Reviews(H_{121},H_{122})**: Both coded as distinct events. We code $Corrections_{tBPM/int}$ if the context of an already explained process part is explicitly changed. In t.BPM sessions this involves re-labeling or (meaningful) repositioning that impacts the process model's meaning. In interviews, explicit revisions of previously stated information is considered a correction. The $Reviews_{tBPM/int}$ are coded if subjects decide to recapitulate their process. This must involve talking about the process as we cannot account possibly silent reviews. This scheme is the same for both experimental tasks.

- **Phases and Problems(H_{123},H_{124})**: As part of the experiment guide, subjects are asked (in separate questions) to name $Phases_{tBPM/int}$ of the process and name $Problems_{tBPM/int}$ that they expect with this process. Using video coding, we count the number of phases and problems reported as events. This applies for both experimental tasks.

Using this coding scheme we operationalize the video hypotheses in the following way:

- H_{114}: Subjects talk more in t.BPM sessions than in interview sessions,
  i.e. $Talk_{tBPM} + UseTalk_{tBPM} > Talk_{int}$.
  H_{014}: Subjects talk equally or less in t.BPM sessions,
  i.e. $Talk_{tBPM} + UseTalk_{tBPM} \leq Talk_{int}$.
- H_{115}: Subjects are more silent in t.BPM sessions than in interviews,
  i.e. $Silence_{t.BPM} > Silence_{int}$
  H_{015}: Subjects are equally or less silent in t.BPM sessions,
  i.e. $Silence_{t.BPM} \leq Silence_{int}$
- H_{121}: Subjects make more reviews in t.BPM sessions than in interviews,
  i.e. $Reviews_{t.BPM} > Reviews_{int}$
  H_{021} Subjects make equally or less reviews in t.BPM sessions,
  i.e. $Reviews_{t.BPM} \leq Reviews_{int}$
- H_{122}: Subjects make more corrections in t.BPM sessions than in interviews,
  i.e. $Corrections_{tBPM} > Corrections_{int}$.
  H_{022} Subjects make equally or less corrections in t.BPM sessions,
  i.e. $Corrections_{tBPM} \leq Corrections_{int}$.
- H_{123}: Subjects report more problems in t.BPM sessions than in interviews.
  i.e. $Problems_{t.BPM} > Problems_{int}$
  H_{023}: Subjects report equally or less problems in t.BPM sessions.
  i.e. $Problems_{t.BPM} \leq Problems_{int}$
- H_{124}: $Subjects report more phases in t.BPM sessions than in interviews.$
  $i.e. Phases_{tBPM} > Phases_{int}$
  H_{024}: $Subjects report equally or less phases in t.BPM sessions.$
  $i.e. Phases_{tBPM} \leq Phases_{int}$

Feedback Hypotheses (H_{133},H_{134}) We operationalize hypotheses related to the feedback test by quantitatively evaluating the amount of mistakes found and comments provided per feedback test. Each test contains seven build-in mistakes. Please note that those mistakes are no hard errors but also cover arguable aspects. A lengthy discussion of this can be found in [54] which is an in depth evaluation of the feedback test results. Subjects are broadly asked to provide feedback. Feedback items are then classified to be either build-in *mistakes* or additional *comments*. Using this coding scheme, we hypothesize:

- H_{133}: Subjects find more mistakes in process models after t.BPM sessions than after interviews.
  i.e. $Mistakes_{tBPM} > Mistakes_{int}$
  H_{033}: Subjects find equally or less mistakes in process models after t.BPM

sessions.
i.e. $Mistakes_{tBPM} \leq Mistakes_{int}$

- H_{134}: Subjects provide more comments to process models after t.BPM sessions than after interviews.
  i.e. $Comments_{tBPM} > Comments_{int}$
  H_{034}: Subjects provide equally or less comments to process models after t.BPM sessions.
  i.e. $Comments_{tBPM} \leq Comments_{int}$

3.6 Variables

The independent variable in this experiment setup is the method used for process elicitation. Subjects do either a structured interview or the same structured interview in the presence of t.BPM, the tangible modeling toolkit. The dependent variables are formed from the data collected during and immediately after a session.

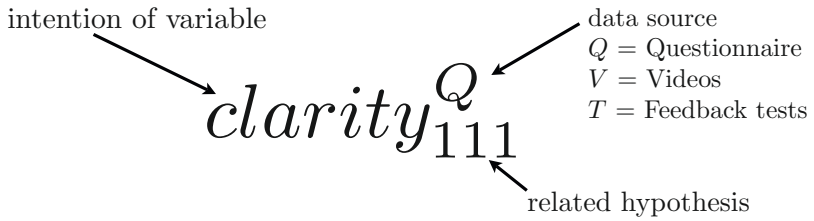


Fig. 4. Notational convention for the dependent variables

We use the following convention to denote the dependent variables:
We use a name to indicate the intention of the variable (e.g. *clarity*). We furthermore denote the number of the related hypothesis that this variable was collected for (e.g. 111). We finally use Q, V or T to indicate how the data for this variable was collected, by questionnaire, video analysis or feedback test.

Each variable is the multiset of the collected values per experimental run, e.g. i.e. $clarity^Q_{111} := \{\{\bigcup_{\forall p \in P} h(p, 6, 15, 8)\}\}$,$|clarity^Q_{111}| = 34$. Accordingly, the other questionnaire related variables are fun^Q_{112}, $motivation^Q_{113}$, $com.solution^Q_{116}$, $insights^Q_{131}$ and $com.feedback^Q_{132}$.

We name the video related hypotheses $talking^V_{114}$, $silence^V_{115}$, $reviews^V_{121}$, $corrections^V_{122}$, $problems^V_{123}$ and $phases^V_{124}$. Due to one missing t.BPM tape only thirtythree variables are in each multiset for video analysis, e.g. $|talking^V_{114}| = 33$. Finally, we define $comments_{133^T}$ and $mistakes_{134^T}$ as the multiset of comments and mistakes collected with the feedback test.

Using this convention we assess the influence of the method on the dependent variables as proposed by the hypotheses. We also use this convention in the

principal component analysis and when testing for further influential factors on the dependent variables.

3.7 Analysis Procedures

Questionnaire data is analyzed by assigning a value [1..5] according to the agreement level per statement as indicated on the Likert scale. Statements that were presented negatively are turned around by calculating '$6 - value$'. Three items in the questionnaire test one hypothesis. We use the average of the three items to test for significant differences between groups.

Video data is analyzed by two independent reviewers. They use the coding scheme specified in Section 3.5, compare their results and (if needed) resolve conflicts by negotiation. The average values (either amount or duration) are used to test for significant differences between groups.

Feedback tests are codified by two experts independently. They classify feedback as a found (build-in) mistake or a comment. Conflicts are resolved by negotiation. The amount of found mistakes or comments is tested for significant differences between groups.

For **statistical evaluation**, we perform and report the following statistical procedures and values:

- Data is tested for normal distribution using Kolmogorov-Smirnov and Shapiro-Wilk test which are requirements to applying the t-test for significance testing.
- A (one-tailed) dependent t-test is used to assess whether the difference between two groups is of statistical significance (p). The acceptance level is at $p<.05$.
- The upper and lower boundaries of the confidence intervals for the mean values are reported. The real mean is in that range with 95 percent probability.
- We use a one-way repeated-measures ANOVA (analysis of variances) to determine the effect caused by the method within individual subjects.

Additionally we perform reliability checks using Cronbach's alpha for the questionnaire items and Cohen's Kappa for inter-rater agreement of the video analysis results. We assess the validity of our hypothesis tree from Figure 3 by using a principal component analysis (PCA). Finally, we use a two-sided t-test to assess factors that potentially have an influence on the performance of the subjects. For this experiment we assess the **potentially influential factors**:

- reported process
  Each subject reports on two different processes, buying a flat screen and running a call for proposals to build a new warehouse. Processes might be unbalanced.
- feedback model
  Each subject gives feedback to two different process models after each treatment, finding a new flat and finding a new job. Models might have different accessibility and difficulty.
- 1st-vs-2nd run
  Each subject goes through the treatment twice. Repetition effects such as learning might influence the performance of the subjects.

- experimenter
  Two experimenters are randomly assigned per subject to run the experiment (with both treatments). Interviewers directly interact with the subjects and might steer results.
- education
  Subjects get slightly different education at school as they are either office clerks or industrial clerks. There might be a difference between these professions.

4 Experiment Execution and Data Collection

The experiment design was executed the week before christmas in 2009. The experimenter team was located in a lecture room at the trade school in Potsdam for one week. Within this week twenty slots were offered to the students by short teasers given in the classes. Students could choose to swap one lecture unit for experiment participation. Seventeen students did take part during the week. Each experiment run started with a short informal warm-up chat and then followed the design as outlined in Section 3.2. One experimenter ran the experiment, the other one operated the cameras and observed the situation to ensure a stable treatment. Figure 5 depicts the two experimental tasks as taped by the cameras. One video taping went wrong, leading to a sample size of sixteen for the video coding hypotheses.

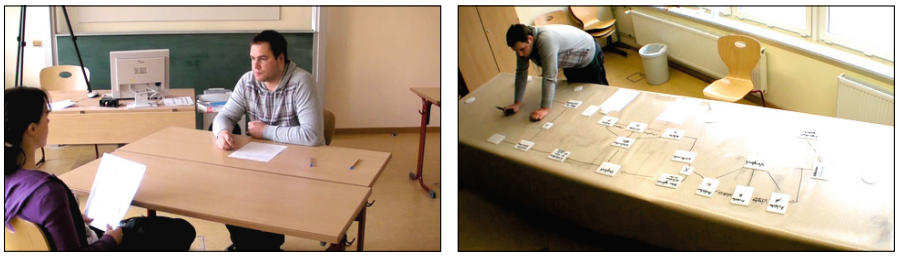


Fig. 5. Fotos from the experiment execution. Subject giving interview (left) and modeling with t.BPM (right). Taped by the video cameras.

We expected to test industrial clerks only. We wanted the most homogenous group possible and we were told, that industrial clerks were in a non-crucial phase of their studies. However, during the week it was not possible to recruit enough industrial clerks. Thus, we opened up the experiment design to both groups, office and industrial clerks. We ended up testing 7 office clerks and 10 industrial clerks. From that we collected thirty-four Feedback tests and thirty-four questionnaires Additionally, more than six hours of video material was collected from taping the experimental tasks. All subjects were at the age of nineteen to twenty-one.

5 Data Analysis

5.1 Descriptive Statistics

From seventeen students, we collected one questionnaire per run. From two runs, this results in two questionnaires each. With eighteen items per questionnaire 612 statements were collected in total for evaluation. Furthermore, the video analysis was conducted based on 6,74 hours of video material. This is based on sixteen t.BPM sessions and seventeen interviews. One t.BPM session taping went wrong. That results in N=16 for all codings that are based on the t.BPM video analysis, such as duration time codings. Table1 describes the video data collected in terms of overall video time and time slices as defined by the formalized hypotheses in Section 3.5. All numbers are depicted in minutes.

Data Type	N	Mean	Min	Max	Standard Deviation	Standard Error
$Video_{tBPM/int}$	16/17	19.52/5.42	10.25/3.53	38.98/9.68	8.22/1.97	2.05/0.48
$Talk_{tBPM/int}$	16/1.7	4.65/3.43	2.62/2.05	10.88/6.60	2.32/1.27	0.58/0.31
$Silence_{tBPM/int}$	16/17	5.54/0.94	1.25/0.27	16.58/2.38	3.86/0.67	0.97/0.16
Use_{tBPM}	16	4.60	1.91	9.86	2.07	0.52
$Use + Talk_{tBPM}$	16	0.64	0	1.37	0.42	0.10
$Rest_{tBPM/int}$	16/17	4.09/1.05	2.56/0.60	6.82/2.43	1.18/0.45	0.29/0.11

Table 1. Descriptive statistics for overall video time and time slices (in minutes) derived from the video analysis

Videos taken during t.BPM sessions took twenty minutes (19.52) on average ranging from ten (10.25) to almost forty minutes (38.98). On the other hand, interviews took about five minutes (5.42) on average ranging from three and a half (3.53) to ten minutes (9.68) at most. The differences in duration of talking and silence are distributed correspondingly.

Data Type	N	Mean	Min	Max	Standard Deviation	Standard Error
$Corrections_{tBPM/int}$	16/17	3/0.29	0/0	6/2	1.9/0.69	0.47/0.17
$Reviews_{tBPM/int}$	16/17	0.81/0.18	0/0	6/1	1.47/0.39	0.37/0.10
$Phases_{tBPM/int}$	16/17	3.56/3.24	2/1	5/5	0.81/0.97	0.2/0.24
$Problems_{tBPM/int}$	16/17	2.63/2.71	0/1	4/4	1.26/1.1	0.31/0.27
$Mistakes_{tBPM/int}$	17/17	2.12/1.94	1/1	5/6	1.45/1.43	0.35/0.35
$Comments_{tBPM}$	17/17	2/2.41	0/1	5/5	1.5/1.18	0.36/0.29

Table 2. Descriptive statistics for events derived from coding videos and mistakes and comments found in feedback test evaluation

Table 2 depicts the amount of events coded during the video analysis as well as the mistakes found and comments given in the feedback tests. While there is a notable difference between t.BPM and interviews in the amount of corrections (3/0.29) and reviews (0.81/0.18). There is only a slight difference between the means of problems, phases, mistakes and comments.

5.2 Data Set Preparation

The data was tested with the Kolmogorov-Smirnov and Shapiro-Wilk test and is normally distributed. No data was excluded from the set. The missing t.BPM video sample was not compensated except for the principal component analysis (see Section 5.3). In that case only, we use the mean value of the t.BPM variables to compensate for the missing 17th data set from the video analysis.

5.3 Measurement Reliability and Validity

According to Kirk [62] the reliability is the extent to which "a measurement procedure yields the same answer however and whenever carried out"([62], p.19) while validity is the "extent to which it gives the correct answer". Transported to our measurement instruments, we assess reliability as the inter-rater agreement for the videos coded and the internal consistency of the questionnaire. Afterwards, we assess the validity of our hypothesis decomposition using a principal component analysis.

Reliability: Cohen's kappa coefficient and Cronbach's alpha In Section 3.5 we propose a coding scheme for the video analysis. The videos were coded by two students independently using VCode [2], a video annotation tool. Deviations in the coding were resolved by negotiation. Using Cohen's kappa coeficient (κ) we measure the inter-coder agreement before the negotiation process. The videos are sliced into intervals of three seconds. Agreement is calculated based on the events seen or not seen in both analysis protocols per interval. The inter-rater agreement over all videos and all coding schemes is $\kappa = .463$. Landis and Koch [63] propose that $0.41 < \kappa < 0.60$ is a moderate agreement level (level five on a seven level scale). Although these levels "are clearly arbitrary" [63] they are frequently used since 1977 to judge κ-values. We interpret our result as an average value indicating suitable coding instructions and reliable (reproducible) video analysis results.

To assess the reliability of the questionnaire we use Cronbach's alpha (α). Each variable measured by the questionnaire is actually split into three statements. Participants agree or disagree with these statements on a five-point Likert scale. We calculate the mean of the three statements to obtain the value for hypothesis testing. Using Cronbach's alpha (α) we measure the degree to which these independent statements coincide. In other words, whether the independent items actually measure the same underlying construct. This is a measure for the internal

[2] http://social.cs.uiuc.edu/projects/vcode.html

consistency of each variable. We calculated $\alpha(clarity^{Q}_{111}) = .687$, $\alpha(fun^{Q}_{112}) = .836$, $\alpha(motivation^{Q}_{113}) = .702$, $\alpha(com.solution^{Q}_{116}) = .911$, $\alpha(insights^{Q}_{131}) = .872$ and $\alpha(com.feedback^{Q}_{132}) = .882$. In the literature [46] >.8 is suggested to be a good value for questionnaires, while >0.7 is still acceptable. We see that α is comfortable for most of our variables. However, $clarity^{Q}_{111}$ and $motivation^{Q}_{113}$ are just at the edge for acceptable reliability. We conclude that reliability of the questionnaire is good for most values. We keep in mind the exceptions for the discussion in Section 6.

Validity: Principal component analysis The principal component analysis [64] reduces our fourteen dependent (possibly correlating) variables into a minimal set of factors. Strongly correlating variables are approximated with one factor called the principal component (pc). If the hypotheses are strictly hierarchically refined, the fourteen variables should be subsumable to three principal components congruent with our hypothesis tree in Figure 3. Using orthogonal (varimax) rotation we identify the five principal components ($pc1..5$) as depicted in table 3.

The amount of principal components and the distribution of variables is not inline with the hypothesis decomposition in Figure 3. The principal component $pc1$ subsumes four variables from the variable set "user dedication" and two variables from the "better feedback" set. A similar effect can be seen in $pc2$ with three variables from two different areas of from the hypothesis tree.

	$pc1$	$pc3$	$pc3$	$pc4$	$pc5$	comment
fun^{Q}_{112}	**.923**	.094	-.022	.060	-.165	all variables measured
$clarity^{Q}_{111}$	**.896**	-.013	-.050	-.040	.050	with the questionnaire
$com.solution^{Q}_{116}$	**.774**	-.152	-.042	.100	.419	
$com.feedback^{Q}_{132}$	**.753**	.123	-.295	.268	.234	
$insights^{Q}_{131}$	**.611**	.030	-.469	-.111	-.415	
$motivation^{Q}_{113}$	**.582**	.413	.262	-.401	-.104	
$talking^{V}_{114}$	-.024	**.877**	.283	.031	-.031	three of six variables
$silence^{V}_{115}$	.053	**.875**	-.182	.288	-.080	measured by video coding
$reviews^{V}_{121}$	.067	**.843**	-.176	.101	.223	
$comments_{133^T}$	-.224	-.008	**.852**	.064	.142	both variables from the
$mistakes_{134^T}$	.027	-.025	**.780**	.053	-.063	feedback test
$corrections^{V}_{122}$	-.029	.149	.185	**.824**	-.106	corrections done
$phases^{V}_{124}$	.122	.139	-.010	**.723**	-.048	and phases reported
$problems^{V}_{123}$	.095	.095	.058	-.176	**.885**	mistakes reported

Table 3. Five principal components identified using oblique (direct oblimin) rotation

We see that $pc1$ subsumes all six variables measured by the questionnaire. These are perceived measures as they are reported by the individuals. In other words, people that report to be motivated, also report to have more fun, more insights et cetera. The individual perception is dominating $pc1$. In $pc2$ variables

from user dedication ($talking^V_{114}$, $silence^V_{115}$) and information quality ($reviews^V_{121}$) are subsumed. This indicates a strong relation between both aspects. Interesting to note, $pc2$ spans three of the six variables measured by video analysis.

The component $pc3$ is composed of the measures collected with the feedback test. In other words, the amount of comments that people give and mistakes that people find in process reviews strongly correlate. Both variables came from the same variable set (better feedback), however they do barely correlate with the other variables in that set ($insights^Q_{131}$, $com.feedback^Q_{132}$). We observe a separation between of perceived performance from objective performance. We investigated this issue in more detail [54] and found that education drives the objective review performance much more than any other value.

The variables in $pc4$ and $pc5$ relate to information quality (H_{12x}). The only variable missing ($reviews^V_{121}$) is in $pc2$. While this seems to be a conforming result at a first glance, we note that H_{12x} is actually distributed over three principal components.

In summary, when matching the principal components with our original hypothesis tree, we find only little coherence. In other words, this instrument does hardly conform our hypothesis decomposition. It is interesting to see, that no principal component spans different instruments. We see a clear separation between the perceived performance (questionnaire) the observed behavior (video analysis) and the measured learning effect (feedback test). This appears to be a strong driver for the distribution of the principal components.

5.4 Hypothesis Testing

The data was tested with the Kolmogorov-Smirnov and Shapiro-Wilk test and is normally distributed. We use a one-sided t-test because we hypothesize a directed effect. Table 4 summarizes the results for the variables tested.

As shown in table 4, there is a significant ($p<.05$) difference for fun^Q_{112}, $talking^V_{114}$, $silence^V_{115}$, $reviews^V_{121}$, $corrections^V_{122}$ and $insights^Q_{131}$ between the groups using t.BPM or interviews. When checking for the confidence intervals we see that $silence^V_{115}$, $corrections^V_{122}$ and $insights^Q_{131}$ have positive ranges. That means with 95 percent probability the true difference between the means of the groups is in a range that does not include zero. Using this standard, we reject the null hypotheses H_{015}, H_{022} and H_{031} as defined in Section 3.5. Although fun^Q_{112}, $talking^V_{114}$ and $reviews^V_{121}$ show significant differences (see table 4), their confidence intervals do include zero. That means, there is a chance that there is actually no effect between the two groups. Thus, we cannot formally reject H_{012}, H_{014} and H_{121} together with all other hypotheses for which we could not find a significant ($p<.05$) difference between the groups. We discuss the result in Section 6.1. There we also take into account the learnings from the influential factors (table 7) and the principal component analysis.

dependent variable	Effect Size t.BPM	interview	Significance	Confidence Intervals lower boundary	upper boundary
$clarity^{Q}_{111}$	3.37	3.49	.304	-0.59	0.36
fun^{Q}_{112}	4.16	3.90	**.046**	-0.05	0.56
$motivation^{Q}_{113}$	4.45	4.37	.225	-0.14	0.29
$talking^{V}_{114}$	4.65	3.49	**.044**	-0.19	2.52
$silence^{V}_{115}$	5.54	0.95	**.000**	**2.63**	**6.54**
$com.solution^{Q}_{116}$	3.31	3.51	.118	-0.53	0.14
$reviews^{V}_{121}$	0.81	0.19	**.033**	-.046	1.30
$corrections^{V}_{122}$	3.00	0.31	**.000**	**1.85**	**3.53**
$problems^{V}_{123}$	2.63	2.81	.327	-1.06	0.69
$phases^{V}_{124}$	3.56	3.19	.094	-0.20	0.95
$insights^{Q}_{131}$	3.75	3.43	**.017**	**0.03**	**0.60**
$com.feedback^{Q}_{132}$	4.14	3.98	.162	-0.17	0.48
$comments^{T}_{133}$	2.00	2.41	.144	-1.21	0.38
$mistakes^{T}_{134}$	2.12	1.94	.191	-0.24	0.59

Table 4. Comparing groups by method using effect sizes, (one-tailed) t-test, and confidence intervals

5.5 Repeated-Measures ANOVA

The analysis of variance (ANOVA) is a family of statistical tests to compare groups in different conditions and explain the variation in a set of dependent variables with the variation from one independent variable. To do that, the data set is partitioned and sums of squares of deviations from the mean value (SS) per group are compared. We use a special case, the repeated-measures ANOVA, to determine the effect of our independent variable (method) within each individual per dependent variableIn other words, to what extend did the method influence the performance of each individual? How much of the performance difference is (un)explained by the method? In Figure 6 we illustrate how our data is partitioned for the repeated-measurement ANOVA. From the overall variability (SS_T), we identify the performance difference within participants (SS_W) and can further distinguish the variation caused by the treatment (SS_M) and the variation not explained by our treatment(SS_R).

The ratio of explained to unexplained variability in our dataset is described by $F = \frac{SS_M}{df_M} / \frac{SS_R}{df_R}$. Where df are the degrees of freedom calculated from the number of different methods (df_M=2-1=1) and the participant number (df_R=17-1=16). The critical ratio $F_{.05}(df_M, df_R)$ is the value to pass before the result is actually significant with an acceptance level of p<.05. For our variables collected in questionnaires and feedback tests $F_{.05}(1, 16) > 4.49$ is a significant result, for the video codings we only have N=16 thus $F_{.05}(1, 15) > 4.54$ is a significant ratio. In table 5 values that are above $F_{.05}$ are highlighted in bold. We also report SS_B, SS_M, SS_R and η^2 (eta squared). The value of $\eta^2 = \frac{SS_M}{SS_W}$ describes the ratio of variation within the subjects that can be explained by the treatment method. It is an effect size measure.

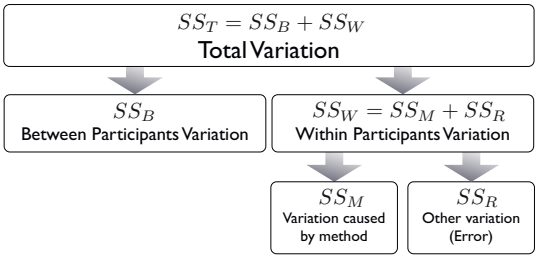


Fig. 6. Data partitioning for the repeated-measuremes ANOVA. Drawing adopted from [46] p.463

dependend Variable	df_R	SS_T	SS_B	SS_M	SS_R	$F_{.05}$	η^2
$clarity^Q_{111}$	16	32.78	25.78	0.12	6.88	0.27	0.02
fun^Q_{112}	16	18.31	15.03	0.55	2.73	3,24	0.17
$motivation^Q_{113}$	16	10.90	9.46	0.05	1.39	0,23	0.04
$talking^V_{114}$	15	116.56	56.92	10.86	48.79	3.34	0.18
$silence^V_{115}$	15	398.55	129.58	167.92	101.05	**24.93**	0.62
$com.solution^Q_{116}$	16	24.68	20.90	0.33	3.45	1.52	0.09
$reviews^V_{121}$	15	38.01	23.00	3.13	11.88	3.95	0.21
$corrections^V_{122}$	15	119.22	42.72	57.78	18.72	**46.3**	0.76
$problems^V_{123}$	15	40.47	19.97	0.28	20.22	0.21	0.01
$phases^V_{124}$	15	25.51	15.50	1.13	8.88	1.9	0.11
$insights^Q_{131}$	16	18.24	14.9	0.84	2.50	**5.36**	0.25
$com.feedback^Q_{132}$	16	19.44	15.99	0.21	3.24	1.03	0.06
$comments^T_{133}$	16	59.56	39.06	1.44	19.06	1.21	0.07
$mistakes^T_{134}$	16	66.98	61.47	0.27	5.24	0.81	0.05

Table 5. ANOVA result table based on df_M=1. Significant $F_{.05}$ ratios marked bold.

From the F ratios we can see that $silence^V_{115}$, $corrections^V_{122}$ and $insights^Q_{131}$ show significant difference due to the method. In those cases the method also accounts for up to seventy-six percent (η=0.76) of the effect within the subjects. Interestingly, in the light of he t-test results from Section 5.4 is that fun^Q_{112}, $reviews^V_{121}$ and $talking^V_{114}$ just missed the significance level.

Important to notice, table 5 also clearly shows that the variation within the subjects are very small when compared with the variation between the subjects. In other words, the differences between the people are much more dramatic then any difference measured within the people, for example cause by treatment method.

In summary, $silence^V_{115}$, $corrections^V_{122}$ and $insights^Q_{131}$ are significantly varying within a subject caused by the treatment method. However, on average the variation between subjects is much bigger.

5.6 Testing potentially influential factors

As described in Section 3.7 we test five factors for their potentially significant influence on the dependent variables. We use each potential factor as an independent variable and assess with a two-sided t-test for significant differences in the mean values collected for the hypotheses. In table 6 we report the significance as found by the t-test.

Important to note, experimenter and education were assessed with an independent t-test because these varied between participants. All other influence factors were assessed with a dependent t-test as they result from a paired set of samples.

	Influence Factors				
	reported process	feedback model	1st-vs-2nd run	experimenter	education
$clarity^{Q}_{111}$	.160	.865	**.001**	.750	**.031**
fun^{Q}_{112}	.532	.901	.091	.618	.1
$motivation^{Q}_{113}$	**.013**	1	.45	.066	.919
$talking^{V}_{114}$	.963	.121	.854	.221	.108
$silence^{V}_{115}$	.738	.407	.996	.448	.316
$com.solution^{Q}_{116}$	.484	1	**.004**	.516	.140
$reviews^{V}_{121}$	.493	.483	.483	.16	.407
$corrections^{V}_{122}$	.700	.177	.939	.119	.660
$problems^{V}_{123}$	.882	.167	**.014**	.662	.344
$phases^{V}_{124}$	1	.388	.669	1	.319
$insights^{Q}_{131}$	.439	1	.439	.984	**.022**
$com.feedback^{Q}_{132}$	1	.626	1	.429	.069
$comments_{133^T}$	.881	.881	.653	.786	**.004**
$mistakes_{134^T}$	.382	.136	.773	.270	**.012**

Table 6. Influential factors tested for their significance (two-tailed t-test)

In table 6 we see that three factors have significant effect on our variables, namely, the reported process, the 1st-vs-2nd run, and education. We investigate those further by reporting effect size and confidence intervals for each significant effect in table 7.

When the process 'call for tenders to build a new warehouse' was used the reported $motivation^{Q}_{113}$ was significantly (p=.013) higher than in runs with the process 'purchase a new flatscreen' (warehouse=4.53, flatscreen=4.29). The 2nd run led to significantly more $clarity^{Q}_{111}$ about the goal (1st=3.1, 2nd=3.77, p=.001), more commitment to the solution ($com.solution^{Q}_{116}$), 1st=3.2, 2nd=3.63, p=.004) and to the awareness of more potential problems in the process ($problems^{V}_{123}$, 1st=2.25, 2nd=3.19, p=.014).

Education had significant impact on four variables. In particular, office clerks reported a more clarity ($clarity^{Q}_{111}$, o-clerks=3.98, i-clerks=3.05, p=.031) and

more new insights into process thinking ($insights^{Q}_{131}$, o-clerks=4.05, i-clerks=3.30, p=.022). Industrial clerks gave significantly more comments ($comments_{133^T}$, o-clerks=2.71, i-clerks=5.60, p=.004) and found more mistakes in the feedback ($mistakes_{134^T}$, o-clerks=2.29, i-clerks=5.30, p=.012).

	Effect Size		Confidence Intervals	
	warehouse	flatscreen	Lower boundary	Upper boundary
$motivation^{Q}_{113}$	4.53	4.29	0.05	0.42
	1. run	2. run		
$clarity^{Q}_{111}$	3.1	3.77	-0.99	-0.34
$com.solution^{Q}_{116}$	3.2	3.63	-0.70	-0.16
$problems^{V}_{123}$	2.25	3.19	-1.65	-0.22
	office clerks	industrial clerks		
$clarity^{Q}_{111}$	3.98	3.05	-1.76	-0.10
$insights^{Q}_{131}$	4.05	3.30	-1.37	-0.12
$comments_{133^T}$	2.71	5.60	1.09	4.68
$mistakes_{134^T}$	2.29	5.30	0.81	5.22

Table 7. Effect sizes and confidence intervals for significantly (p<.05) influential factors on the dataset.

6 Result Discussion

To discuss the results we touch each hypothesis in Section 6.1. Section 6.2 discusses the implications of influential factors as identified in Section 5.6. Afterwards we discuss validity threats in Section 6.4 and the generalizability of our findings in Section 6.5. We conclude this section with a summary of the discussion in Section 6.7.

6.1 Hypotheses discussed

Out of fourteen hypotheses, six showed significant differences between t.BPM and interviews. However, three of them have a critical confidence interval. We interpret each hypothesis in the light of the t-tests, the repeated-measures ANOVA, the principal component analysis and potentially influential factors on the data set. This section closes with a summary of the conclusions drawn from this discussion.

- H_{111}: On average, participants report no significantly (p=.304) clearer understanding of the goal due to t.BPM (t.BPM=3.37,int=3.49). Thus we do not reject H_{011}. In contrast to our deliberations in the hypothesis creation, t.BPM does not automatically imply that people understand what is the expected outcome of the session is.

Interestingly, repetition and education had significant influence on $clarity^{Q}_{111}$. On average, subjects reported a significantly (p=.001) clearer goal understanding in the second experimental task (1st=3.1,2nd=3.77, see table 7). We interpret this as the learning effect in our repeated measurement design. Participants in the second run had already experienced the experiment situation and therefore more clarity. We also found a significant (p=.001) difference between subjects with different education (o-clerks=3.98, i-clerks=3.05, see 7). We attribute this to a tendency for office clerks to report a more positive self image, see Section 6.2 for details.

- H_{112}: On average, participants report significantly (p=.046) more fun in t.BPM sessions (t.BPM=4.16,int=3.90). However, the confidence interval (lb=-0.05,ub=0.56) is critical as it includes zero. Thus, we do not formally reject H_{012}. A slightly larger sample size would probably have changed this fact. We draw this also from the repeated-measures ANOVA in which fun^{Q}_{112} slightly missed the critical ratio ($3.24 < F_{0.5}(1, 16) = 4.49$). Interestingly, fun^{Q}_{112} is a key driver for principal component $pc1$. We discuss the implications in Section 6.3.
- H_{113}: On average, participants report no significantly (p=.225) higher motivation due to t.BPM (t.BPM=4.45,int=4.37). Thus we do not reject H_{013}. We interpret this result as a ceiling effect. On a five point Likert scale, people scored a 4.41 on average. That might result from the incentives (chocolate + cinema vouchers + off from school) or simply the fact that people volunteered to participate. In any case, a significant difference cannot be found in this small sample set, although there is a positive trend towards t.BPM. Interestingly, the type of process used in the experimental task significantly (p=.013) influences $motivation^{Q}_{113}$ (warehouse=4.53,flatscreen=4.29 see table 7). We can only assume that the warehouse process was more realistic and challenging to work with.
- H_{114}: On average, participants talk significantly (p=.044) more in t.BPM sessions (t.BPM=4.65min,int=3.49min). However, the confidence interval (lb=-0.19,ub=2.52) is critical as it includes zero. Analogue to H_{112} we do not formally reject the null hypothesis (H_{014}) but we assume that this would be possible with a slightly larger sample set. While the average difference between t.BPM and interviews is considerable (1.16min), only twelve seconds (-0.19min=11.63sec, see table4) are missing at the lower boundary of the confidence interval. Again, $talking^{Q}_{114}$ also just misses the critical ratio ($3.34 < F_{0.5}(1, 16) = 4.49$) in the repeated-measures ANOVA.
- H_{115}. On average, participants spent significantly (p=.000) more time silent in t.BPM sessions (t.BPM=5.54min,int=0.95min). Since the confidence interval (lb=2.63,up=6.54) is also positive, we reject H_{015}. This finding is also supported by the repeated-measures ANOVA ($F_{0.5}(1, 15) = 24.93$) in which the method can explain 62 percent of the effect within the subjects ($\eta^2 = 0.62$). We conclude that the presence of t.BPM makes people spent more time silent. Although we can only judge on the observed behavior, we interpret the silent time as time taken to think about the process. We conclude that t.BPM affords people to

think deeply in elicitation sessions in contrast to interviews in which talking is the purpose of the session.

- H_{116}: On average, participants report no significantly (p=.118) higher commitment to the solution due to t.BPM (t.BPM=3.31,int=3.51). Thus we do not reject H_{016}. The tendency even points into the opposite direction. In the repeated-measures ANOVA we see that most of the effect is between subjects (SS_B=20.90,SS_W=3.78) indicating that the commitment is strongly depending on the people rather than the method. However, just like $clarity^Q_{111}$, the commitment to the solution is significantly (p=.004) higher for the second experimental task (1st=3.2, 2nd=3.66, see table 7). We also see in table 3 that both variables are part of the same principal component ($pc1$) which means $com.solution^Q_{116}$ strongly correlates with $motivation^Q_{113}$. We conclude that method alone does not make people be more committed to their solution. Instead, repetition leads to more clarity and also more confidence about the produced results.
- H_{121}: On average, participants do significantly (p=.033) more reviews in t.BPM sessions (t.BPM=0.81,int=0.19). However, the confidence interval (lb=-0.46,ub=1.30) is critical as it includes zero. Thus, we do not formally reject the null hypothesis. Again, the repeated-measures ANOVA confirms this result with the critical ratio just missed ($3.95 < F_{0.5}(1, 15) = 4.54$) by $reviews^V_{121}$. During experiment execution, experimenters were not supposed to trigger reviews, e.g. by asking for them. Furthermore, reviews that people did not articulate were not counted as $reviews^V_{121}$ but as $silence^V_{115}$. Thus, only intrinsically started verbal reviews are considered here. Since we see a significant difference and with comfortable confidence intervals (lb=2.63, ub=6.54) for H_{115}, it might be that some reviews were done silently and therefore are not included here. Nonetheless, we can only conclude that there is a significant difference in reviews. However, we cannot conclusively state a difference for both groups with 95 percent probability as we can for $silence^V_{115}$. A larger sample size or a different coding scheme would probably change this.
- H_{122}: On average, participants do significantly (p=.000) more corrections in t.BPM sessions (t.BPM=3.00,int=0.31). Since the confidence interval (lb=1.85,up=3.53) is also positive, we reject H_{022}. The result from the t-tests are also confirmed by the ANOVA results($F(1,15)$= 46.3, η^2=0.76). Our result confirms the relevance of mapped representations to reduce cognitive load and enable for instant feedback [52, 53, 55]. We conclude that the use of t.BPM leads to corrections which we see as iterations of the process model. In our opinion it also supports H_{121} because corrections to previously stated information typically require a review of the information first.
- H_{123} & H_{124}: On average, participants did not state significantly (p=.327) more problems in the process in t.BPM sessions (t.BPM=2.63,int=2.81). Likewise, participants did not state significantly (p=.094) more phases in the process in t.BPM sessions (t.BPM=3.56,int=3.19). In both cases, we simply were wrong with our hypotheses. We assumed that the mapping effect [53, 65] would also lead people to report more fine grained about problems and phases in their

process. It turns out participants report about three phases and problems either way. In Section 5.6 we identified repetition again as an influential factor. On average, participants reported significantly (p=.014) more problems in the second experimental task (1st=2.25, 2nd=3.19, see table 7).

- H_{131}: On average, participants did report significantly (p=.017) more insights into process thinking due to t.BPM (t.BPM=3.75,int=3.43). Since the confidence intervals (lb=0.03, up=0.60) are also positive, we reject H_{031}. This is also confirmed by result from the repeated-measures ANOVA ($F(1,16)$=5.36,η^2=0.25). This hands-on learning effect was hypothesized, yet we are cautious. When looking at $comments^T_{133}$ and $mistakes^T_{134}$ we see, that reported insights does not lead to better feedback on process models. In other words, there is a mismatch between the perceived task performance and the measured task performance. We attribute this to a flawed self-perception. As an example, on average, office clerks report significantly (p=.022) more insights (o-clerks=4.05,i-clerks=3.05) than industrial clerks but score worse with objective measures (see $comments^T_{133}$ & $mistakes^T_{113}$ in table 7). We interpret this as a perceived higher learning with t.BPM for office clerks but we are also aware that office clerks tend to report a more positive self-image. For more details about the differences between the groups, see Section 6.2 and [54].
- H_{132} On average, participants did not report significantly (p=.162) more commitment to the feedback due to t.BPM (t.BPM=4.14,int=3.98). Thus, we do not reject H_{032}. With a much larger sample set, the significance might reach acceptance level (p<.05), but also the effect size is not very big. Thus, we conclude that t.BPM may not have impact on the commitment to better feedback.
- H_{133} and H_{134}: On average, participants did not provide significantly (p=.144) more comments to process models after having done t.BPM (t.BPM=2.00, int=2.14). Likewise, participants did not find significantly (p=.191) more mistakes in process models after having done t.BPM (t.BPM=2.12, int=1.92). Thus we do not reject neither H_{033} nor H_{034}. The perceived learning with t.BPM that was reported in $insights^Q_{131}$ is not reflected in the feedback results. We conclude that there is a gap between the perceived learning and the measured effect. In the repeated-measures ANOVA we see that the variation between people (SS^{133}_B=39.06,SS^{134}_B=66.98) is much bigger than the variation within people(SS^{133}_W=20.5,SS^{134}_W=5.51). Transported to reality it means that consultants cannot expect better reviews by using t.BPM in elicitation sessions. Instead, the people's background matters. In Section 5.6 we identified education to have significant influence on $comments^T_{133}$ (p=.004) and $mistakes^T_{134}$ (p=.012). In [54] we investigate the feedback test results and this interrelation in deep detail. We pick up the discussion about the influence of education here in Section 6.2.

In summary, we reject H_{015}, H_{022} and H_{031} as statistics show significant difference with comfortable confidence intervals. We argue that H_{112}, H_{114} and H_{121} just slightly failed statistical relevance but could be accepted with a larger sample set. The other hypotheses did not hold. However, we assume a ceiling

effect for $motivation^{Q}_{113}$ and we observed that $clarity^{Q}_{111}$ and $com.solution^{Q}_{116}$ significantly increase as people get more experienced (2nd run). Likewise people reported more $problems^{V}_{123}$ in the second experimental task. Otherwise, $problems^{V}_{123}$ and $phases^{V}_{124}$ reported were not significantly influenced by any variable including t.BPM usage. Finally, we see that t.BPM does not increase feedback ($comments^{T}_{133}$,$mistakes^{T}_{134}$) on process models or even only the commitment to give better feedback ($com.feedback^{Q}_{132}$). However, other factors do.

6.2 Influential Factors discussed

We noted in Section 5.6 that **education** is the most influential factor in our data set with four variables significantly concerned, see table 6. Interestingly, there is a counter effect between the perceived performance measures ($clarity^{Q}_{111}$ & $insights^{Q}_{131}$) and the objective performance measures ($comments^{T}_{133}$ & $mistakes^{T}_{134}$). In other words, on average, office clerks report to have significantly (p=.031) better goal understanding (o-clerks=3.98,i-clerks=3.05) and significantly (p=.022) more new insights into process thinking (o-clerks=4.05,i-clerks=3.30). But the same group finds significantly (p=.012) less mistakes in the process models (o-clerks=2.29, i-clerks=5.30) and gives significantly (p=.004) less comments (o-clerks=2.71, i-clerks=5.60). We were astonished by the effect sizes and checked the mean differences between the two groups for all dependent variables. It turns out that office clerks tend to score higher on the Likert scale which results in more positive answers for all questionnaire variables. However, the effect was not significant except for $clarity^{Q}_{111}$ and $insights^{Q}_{131}$. We think that those two variables in particular illustrate the gap between the perceived performance and the measured performance of a person (here $comments^{T}_{133}$ & $comments^{T}_{134}$). This is an important take away from this experiment.

We also conducted a post-experiment interview with the principal of the trade school. When asked about the differences between the two professions, we were told that the performance in school and the job positions of those people also differs. As we explained in Section 3.4 office clerks typically work as office managers and conduct supporting activities within a department. On the contrary, industrial clerks are key personnel doing planing, execution and controlling of operational business activities. We note this, but do not see a harm to our experimental results. Both groups are potential users of t.BPM and significant influence on the data set is limited to four variables of which only $insights^{Q}_{131}$ is also significant (p=.017) with respect to method (t.BPM=3.75,int=3.43). We conclude that office clerks (think to) learn more in t.BPM sessions. In [54] we elaborate in detail on the performance of the groups in the context of the feedback test.

The second biggest significant influence is attributed to the **learning effect**. The variables $clarity^{Q}_{111}$, $com.solution^{Q}_{116}$ and problems have significant influence (see table 6) with considerable effect sizes (see table 7). We conclude that clarity of the goal raises with repetition, as proposed by the literature [66]. We note that commitment to the solution increases ($com.solution^{Q}_{116}$) as well and trace this back to more confidence that people build as they repeat a task.

For $problems^{V}_{123}$, we argue that people simply take more time to think about them more deeply in the second treatment because they learned already in the first run that this would be the last part of the experimental task. As shown in Figure 11 in Appendix A, asking for problems is the very last step in the experimental treatment. Thus, subjects in the second run might have taken more time to think about this more deeply and thus came up with more problems (1st=2.25, 2nd=3.19).

Finally, the variable $motivation^{Q}_{113}$ is significantly (p=.001) higher for the process 'call for tenders to build a new warehouse' than for the process 'purchase a new flatscreen' (warehouse=4.53, flatscreen=4.29). We do not have a good explanation for this. It might be that the flatscreen-process was less realistic or less challenging to investigate and therefore less interesting and motivating.

In summary, we see that education is significantly influential in our data set. We elaborate this for the feedback tests in a separate publication [54]. Here we note that office clerks perform worse but tend to perceive their performance more positive than industrial clerks. We also observed some learning effects which could have been expected. Unclear remains, why one process motivated participants more than the other. We consider none of the influences to be harmful for our findings.

6.3 Principal components discussed

We recognize that $pc1$ is completely made up of the variables collected through the questionnaire, see table 3. We interpret this as a positive result as it shows that all perceived measures strongly correlate.

Similarly, we are not surprised about $pc2$ covering $talking^{V}_{113}$, $silence^{V}_{115}$ and $reviews^{V}_{121}$. We interpret $pc2$ as the collection of variables that indicate the profoundness with which people expose themselves to the experimental task. We did not limit time. Thus, subjects decided to spent time talking and thinking about the process. It is only consequent that those people also more often decide to do more $reviews^{V}_{121}$ of the process.

In $pc3$ the feedback test variables are collected. This is confirming previous work [54] in which we found strong individual differences in feedback performance. It also has a very practical reason: We asked people loosely to provide feedback and classified the feedback as found $mistakes^{T}_{134}$ or additional $comments^{T}_{133}$. Naturally, people providing more feedback end up with more items in each category. Thus, we see a strong correlation of these two variables.

We interpret $pc4$ and $pc5$ as the remaining variables that are left over. We cannot think of a good reason why the amount of $corrections^{V}_{122}$ that a person applies to their process story should strongly correlate with the amount of $phases^{V}_{124}$ that she names. We would have expected $phases^{V}_{124}$ an $problems^{V}_{123}$ to be in one principal component, as both are collected very similarly through the last two questions asked in the experimental task (see Figure 11 in Appendix A). Nevertheless, distribution of these three variables on the last two principal components is not too surprising.

In summary, the principal component analysis reveals the strong correlation of the perceived measures ($pc1$). It also indicates three variables that measure the profoundness ($pc2$) with which subjects work in the experimental task. Finally, it confirms the independence of the feedback test variables ($pc3$). The building of $pc1-3$ is a good support for the reliability of our measurement instruments. In other words, it indicates the consistency of our measurements with their intention.

6.4 Validity threats

Some issues of **internal validity** were addressed in this experiment by design. In particular, we use two processes, two feedback models and two experimenters assigned in random order. In Section 5.6 we assess potentially confounding variable for their influence. We found learning effects due to the repeated measurements design. For example, participants report a clearer goal understanding for the second experimental treatment. As discussed in Section 6.2 we do not see a harm for our results. A dominant influential factor was the heterogeneity of the group, consisting of office and industrial clerks.

While group heterogeneity is a threat to the internal validity, it also increases the **external validity** as both groups represent the population that we would like to generalize upon. In general, we did our best to mimic a field situation in which consultants do structured interviews and collect feedback in the model afterwards. Choosing domain processes rather then artificial graphs was important to keep generalizability. However, process content such as 'moving to a new flat' is a personal, not a business process.

For all **measurement instruments**, one might argue that they are not well chosen and tested. This is apparent from the small portion of hypotheses that were accepted and the principal component analysis which did not confirm our tree-like hypothesis decomposition. We decomposed the hypotheses to represent the expected benefits in the field. The operationalization of hypotheses was tested in one pre-study with ten computer science students. Adjustments were made afterwards. To ensure quality standards for data evaluation, we used two independent coders for the video analysis, two experts for the feedback test evaluation, and we have split each questionnaire variable into three items, one poled negatively. Finally, we provide all experimental material in the Appendix A for interested readers.

6.5 Generalizability of findings

As mentioned before, we think the findings about t.BPM can be generalized from the sample group to the general population. All participants are affiliated with companies and represent exactly the group we tend to address with the t.BPM tool. Moreover, we think we can generalize the effect of tangible prototyping in contrast to pure talking.

We have observed that people spent significantly more time thinking and talking if t.BPM, an external visualization, is present. The same treatment also

led to significantly more corrections. We think that the affordance of an external visualization in addition to the discussed knowledge is nothing specific to t.BPM. Other visual mappings have been reported to provoke the same effect [55].

The aspect of tangibility enables novices to easily work with the representation and express their knowledge. This leads to deeper involvement and a stronger learning effect (here insights) through hands-on experience also with other knowledge representations, not only processes or t.BPM.

6.6 Lessons learned

Running the experiment enabled us to learn by doing. If we had to start over again, we would probably put even more effort into instrument validation, such as the questionnaire. However, we also learned that people may report a wrong self-image that has to be validated with more objective measures. Yet, we would probably not have too many video-based hypotheses. The video analysis phase consumed most effort of the overall experiment evaluation.

Besides, the compact on-site experiment was a good idea. Instead of spreading it out over various weeks with changing conditions, we could collect the data in a compact week with a stable setup. Moreover, the two experimenters which reviewed each others work did ensure a stable setup. Of course, we should have split the experiment into many experiments with less complexity each. However, we do not get this chance very often.

6.7 Summarizing the discussion

Out of our fourteen hypotheses, we found six to show significant differences between the groups. Three of them (H_{112},H_{114},H_{121}) showed critical confidence intervals. Thus, we formally accept only H_{115},H_{122},H_{131} by rejecting the according null hypotheses (H_{015},H_{022},H_{031}). This finding was confirmed by a repeated-measures ANOVA. We discuss in Section 6.1 that the hypotheses with critical confidence intervals (H_{112},H_{114},H_{121}) could probably be accepted with a slightly bigger sample size.

In detail, we found that people take more time to talk (H_{114}) and think (H_{115}) about their process. They more often review (H_{121}) and correct (H_{122}) their processes. Finally, they report to have more fun (H_{112}) and more new insights (H_{131}) into process modeling.

By investigating influential factors in Section 6.2 we have identified repetition to be crucial for clarity of the goal, the commitment to the solution and the problem awareness. This analysis also revealed a mismatch between perceived and objective performance of individuals. While office-clerks score badly in the process feedback task, they tend to report a more positive self-image. The principal component analysis in Section 6.3 underpins this. It subsumes all perceived measures in one big component ($pc1$) and groups the feedback performance metrics in another principal component ($pc3$).

The significant influence of education on, e.g. feedback performance, points out the relevance of the people that are chosen for the task. The results from

the repeated-measures ANOVA very well support this conclusion. The difference between individuals in each variable are dramatically higher than any difference caused by the treatment. In other words, it is about the people much more than about the treatment.

As discussed in Section 6.4 we addressed validity threats in the design and the evaluation of the experiment. We suggest that our findings are not limited to t.BPM but can be generalized to tangible knowledge mapping approaches as discussed in Section 6.5.

7 Conclusion & Future Work

This paper reports on a controlled experiment which was conducted with 17 student clerks at the trade school. We investigated the process elicitation method as an independent variable. Subjects did structured interviews and t.BPM in a repeated measurement design. We hypothesize that t.BPM enables more efficient process elicitation which is broken down to fourteen operationalized hypotheses. These are evaluated by questionnaires, feedback tests and video analysis. Six hypotheses showed significant differences between the groups according to method. We conclude that

- t.BPM creates a different working mode. I.e. people talk more and think more about their process.
- t.BPM fosters instant feedback. I.e. people review their process more often and also apply more corrections during the elicitation session.
- t.BPM is fun to learn with. I.e. people report to have more fun and more new insights into process modeling.

Besides dependent t-tests we also used a repeated-measures ANOVA, explored the data with a principal component analysis, and tested potentially influential factors for their significance. Interrelating these result we conclude that,

- People matter. I.e. there is dramatically more effect between people than within people due to method.
- Repetition matters. I.e. second round showed more clarity and commitment.
- Measures matter. I.e. weaker task performers reported a more positive self-image.

We are aware that this study is limited by the small sample size (N=17) and our measurement instruments. The actual experiment material is appended to this paper to enable readers to make up their own mind about it. We think the findings can be generalized for tangible mapping techniques for which t.BPM is one instance. Furthermore, we fame our learnings as recommendation to practitioners in similar situations:

- Make it tangible. I.e. use an external representation that everybody can touch and interact with.

- Make a warm-up game. I.e. starting with an artificial matter rather than the actual one provides more clarity for the actual task. Ideas for process games can be found in e.g. [19].
- Get the right people. I.e. the performance variation between people is much bigger than any variation caused by treatment.

In future work, a similar study might re-enforce the findings by repeating measures with a larger sample set. We propose to further limit the group for a more homogeneous sample population. For our research, we decide to move on to further develop t.BPM as a method for elicitation. The next step is to take the idea to the field and create a guidance to work out the processes with t.BPM. The findings from this experiment, are the first building block for t.BPM as a process elicitation technique.

Acknowledgements

We are grateful to the students that supported this work. First and for most, Karin Telschow. She helped setting up, running and evaluating this experiment. Likewise, Markus Güntert helped to setup and run the experiment. Finally, we thank Carlotta Mayolo for her support in the video analysis phase.

References

1. Burlton, R.: Business process management. Sams (2001)
2. van der Aalst, W., Hofstede, A., Weske, M.: Business process management: A survey. Lecture Notes in Computer Science **2678** (2003) 1–12
3. Davenport, T.: Process innovation: reengineering work through information technology. Harvard Business School Pr (1993)
4. Hammer, M., Champy, J.: Reengineering the corporation: A manifesto for business revolution. Collins Business (2003)
5. Boehm, B.: Software engineering economics. Prentice Hall (1981)
6. Zave, P., Jackson, M.: Four dark corners of requirements engineering. ACM Transactions on Software Engineering and Methodology (TOSEM) **6**(1) (1997) 1–30
7. Byrd, T., Cossick, K., Zmud, R.: A synthesis of research on requirements analysis and knowledge acquisition techniques. MIS Quarterly (1992) 117–138
8. Davis, A., Dieste, O., Hickey, A., Juristo, N., Moreno, A.: Effectiveness of requirements elicitation techniques: Empirical results derived from a systematic review. In: 14th IEEE International Conference Requirements Engineering. (2006) 179–188
9. Scheer, A.: ARIS-business process modeling. Springer Verlag (2000)
10. OMG: Business Process Modeling Notation (BPMN) 1.2. (January 2009)
11. van der Aalst, W.: Making work flow: On the application of petri nets to business process management. Lecture notes in computer science (2002) 1–22
12. Wohed, P., van der Aalst, W., Dumas, M., Hofstede, A., Russell, N.: On the suitability of bpmn for business process modelling. Lecture Notes in Computer Science **4102** (2006) 161

13. van Dongen, B., van der Aalst, W., Verbeek, H.: Verification of EPCs: Using reduction rules and Petri nets. In: CAISE. Volume 3520., Springer (2005) 372–386
14. Recker, J., Dreiling, A.: Does it matter which process modelling language we teach or use? an experimental study on understanding process modelling languages without formal education. In: 18th Australasian Conference on Information Systems, Toowoomba, Australia, The University of Southern Queensland, Citeseer (2007) 356–366
15. Holschke, O., Rake, J., Levina, O.: Granularity as a Cognitive Factor in the Effectiveness of Business Process Model Reuse. In: Proceedings of the 7th International Conference on Business Process Management, Springer (2009) 260
16. Mendling, J., Reijers, H., van der Aalst, W.: Seven process modeling guidelines (7pmg). Information and Software Technology **52**(2) (2010) 127–136
17. Recker, J., Rosemann, M.: The measurement of perceived ontological deficiencies of conceptual modeling grammars. Data & Knowledge Engineering (2010)
18. Sedera, W., Gable, G., Rosemann, M., Smyth, R.: A success model for business process modeling: findings from a multiple case study. In: Proceedings of The 8th Pacific-Asia Conference on Information Systems, Shanghai, China. (2004)
19. Rittgen, P.: Success factors of e-collaboration in business process modeling. In Pernici, B., ed.: Conference on Advanced Information Systems Engineering (CAISE 2010). (2010) 24–37
20. Recker, J.: Continued use of process modeling grammars : the impact of individual difference factors. European Journal of Information Systems **19** (2010) 76–92
21. Rittgen, P.: Negotiating Models. LECTURE NOTES IN COMPUTER SCIENCE **4495** (2007) 561
22. Susskind, L.: Arguing, Bargaining and Getting Agreement. Oxford Handbook of Public Policy (2006)
23. Rittgen, P.: Collaborative Modelling Architecture (COMA), http://www.coma.nu/COMA-Handbook.pdf. (2008)
24. Hevner, A., March, S., Park, J., Ram, S.: Design science in information systems research. MIS Quarterly **28**(1) (2004) 75–106
25. Lewin, K.: Action research and minority problems. Journal of social issues **2**(4) (1946) 34–46
26. Rittgen, P.: Collaborative Modeling–A Design Science Approach. In: Proceedings of the 42nd Hawaii International Conference on System Sciences (HICSS-42), Waikoloa, Big Island, Hawaii, USA, January. (2009) 5–8
27. Weidlich, M., Zugal, S., Fahland, D., Weber, B., Reijers, H., Mendling, J.: The impact of sequential and circumstantial changes on process models. In: ER-POIS workshop at CAISE, Springer (2010) 43–54
28. Weber, B., Pinggera, J., Zugal, S., Wild, W.: Handling events during business process execution: An empirical test. In: ER-POIS workshop at CAISE. (2010) 19–30
29. Ishii, H., Ullmer, B.: Tangible bits: towards seamless interfaces between people, bits and atoms. In: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM (1997) 234–241
30. van den Hoven, E., Frens, J., Aliakseyeu, D., Martens, J.B., Overbeeke, K., Peters, P.: Design research & tangible interaction. In: TEI '07: Proceedings of the 1st international conference on Tangible and embedded interaction, New York, NY, USA, ACM (2007) 109–115
31. Laurel, B.: Design research: Methods and perspectives. The MIT press (2003)
32. Buxton, W., service, S.O.: Sketching user experiences: getting the design right and the right design. Morgan Kaufmann (2007)

33. Leifer, L.J., Meinel, C.: The Philosophy behind Design. Springer Verlag (2010)
34. Boujut, J., Blanco, E.: Intermediary objects as a means to foster co-operation in engineering design. Computer Supported Cooperative Work (CSCW) **12**(2) (2003) 205–219
35. Gibson, J.: The theory of affordances. Perceiving, acting and knowing: toward an ecological psychology (1977) 67–82
36. Kettinger, W., Teng, J., Guha, S.: Business Process Change: A Study of Methodologies, Techniques, and Tools. MIS Quarterly **21** (1997) 55–80
37. Stirna, J., Persson, A., Sandkuhl, K.: Participative Enterprise Modeling: Experiences and Recommendations. Lecture Notes in Computer Science **4495** (2007) 546
38. Spitz, P.: Quality and the Reengineering Imperative. The chemical industry at the millenium: maturity, restructuring, and globalization (2003) 145
39. Fahrwinkel, U.: Methoden zur Modellierung und Analyse von Geschäftsprozessen zur Unterstützung des Business Process Reengineering. Dissertation, Fakultät für Maschinenbau, Universität Paderborn (1995) EUR 50,00, ISBN 3- 931466-00-0.
40. Nixdorf Institute, H.: OMEGA: Object-Oriented Method Strategic Redesign of Business Processes. In: Changing the ways we work: shaping the ICT-solutions for the next century: proceedings of the Conference on Integration in Manufacturing, Göteborg, Sweden, 6-8 October 1998. (1998) 381
41. Gausemeier, J., Plass, C., Wenzelmann, C.: Zukunftsorientierte Unternehmensgestaltung–Strategien, Geschäftsprozesse und IT-Systeme für die Produktion von morgen. Hanser Fachbuch (2009)
42. Creswell, J.: Research design: Qualitative, quantitative, and mixed methods approaches. Sage Pubns (2008)
43. Wohlin, C., Runeson, P., Höst, M.: Experimentation in software engineering: an introduction. Springer Netherlands (2000)
44. Jedlitschka, A., Ciolkowski, M., Pfahl, D.: Reporting experiments in software engineering. Guide to advanced empirical software engineering (2008) 201–228
45. Harris, P.: Designing and reporting experiments in psychology. Open University Press (2003)
46. Field, A.: Discovering statistics using SPSS. SAGE publications Ltd (2009)
47. Stevenson, A.: Oxford Dictionary of English. Volume 24. Oxford University Press (2010)
48. Xie, L., Antle, A.N., Motamedi, N.: Are tangibles more fun?: comparing children's enjoyment and engagement using physical, graphical and tangible user interfaces. In: TEI '08: Proceedings of the 2nd international conference on Tangible and embedded interaction, New York, NY, USA, ACM (2008) 191–198
49. Schaufeli, W., Martinez, I., Pinto, A., Salanova, M., Bakker, A.: Burnout and engagement in university students: A cross-national study. Journal of Cross-Cultural Psychology **33**(5) (2002) 464
50. Schaufeli, W., Salanova, M., González-Romá, V., Bakker, A.: The measurement of engagement and burnout: A two sample confirmatory factor analytic approach. Journal of Happiness Studies **3**(1) (2002) 71–92
51. Sweller, J., Chandler, P.: Evidence for cognitive load theory. Cognition and Instruction **8**(4) (1991) 351–362
52. Miller, G.: The magical number seven, plus or minus two. Psychological review **63** (1956) 81–97
53. Zhang, J.: The nature of external representations in problem solving. Cognitive science **21**(2) (1997) 179–217

54. Grosskopf, A., Weske, M.: On business process model reviews. In: In workshop proceedings of ER-POIS: Empirical Research on Process Oriented Information Systems affiliated to CAiSE10, Springer (2010) 31–42
55. Schneider, K.: Generating Fast Feedback in Requirements Elicitation. Lecture Notes in Computer Science **4542** (2007) 160
56. Bruner, J.: The act of discovery. Harvard Educational Review **31**(1) (1961) 21–32
57. Mayer, R.: Models for understanding. Review of educational research **59**(1) (1989) 43
58. Melcher, J., Mendling, J., Reijers, H., Seese, D.: On measuring the understandability of process models (experimental results). In: Proceedings of the 1st International Workshop on Empirical Research in Business Process Management (ER-BPM). (2009)
59. Laue, R., Gadatsch, A.: Measuring the understandability of business process models - are we asking the right questions? In: Proceedings of the 6th International Workshop on Business Process Design (BPD 2010). (2010)
60. Greenwald, A.: Within-subjects designs: To use or not to use. Psychological Bulletin **83**(2) (1976) 314–320
61. Cooper, D., Schindler, P.: Business Research Methods. 10 edn. McGraw-Hill Higher Education (2008)
62. Kirk, J., Miller, M.: Reliability and validity in qualitative research. Sage Publications, Inc (1986)
63. Landis, J., Koch, G.: The measurement of observer agreement for categorical data. Biometrics **33**(1) (1977) 159–174
64. Jolliffe, I.: Principal component analysis. Springer Verlag (1986)
65. Tversky, B.: Some ways that maps and diagrams communicate. Lecture Notes in Computer Science **1849** (2000) 72–79
66. Beckman, S., Barry, M.: Innovation as a learning process: Embedded design thinking. Harvard Business Publishing (Nov 2007)

A Experimental Material

Der Begriff Geschäftsprozessmanagement
Geschäftsprozessmanagement ist ein ganzheitlicher Ansatz zur Optimierung der Abläufe im Unternehmen. Diese Abläufe können das Eintreffen neuer Waren oder das Bezahlen von Mitarbeitern sein. Diese Methodik ist nicht auf Unternehmensbereiche begrenzt sondern versucht die Wertschöpfung im Unternehmen nachzuvollziehen. Das Ziel ist das Unternehmen effektiver zu machen, indem man die Abläufe besser versteht und optimiert. Im unten stehenden Bild werden die Phasen des Prozesslebenszyklus beschrieben.

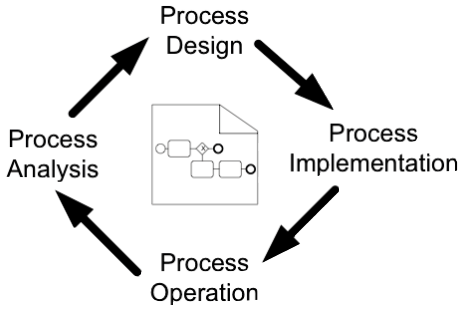


Bild1: Geschäftsprozesslebenszyklus
(Process Management Lifecycle)

Im ersten Schritt werden die Prozesse analysiert (Process Analysis). Die Frage ist „Wer macht was, wann, wie und womit?“. Man bezeichnet dies als den Ist-Prozess. Es ist die aktuell gelebte Realität. Im zweiten Schritt wird der neue, optimierte Prozess entworfen. Es ist der Soll-Prozess. Hier wird der gewünschte Zustand beschrieben. Dabei wird der Prozess mit allen Beteiligten diskutiert um mögliche Probleme oder Optimierungsmöglichkeiten zu identifizieren.
Wenn Einigkeit über den gewünschten Prozess besteht muss er umgesetzt werden (Process Implementation). Bei der Umsetzung durch Software kann dies Programmieraufwand bedeuten. In jedem Fall jedoch müssen Mitarbeiter geschult und auf den neuen Prozessablauf eingeschworen werden. Wenn der Prozess dann wirklich gelebt wird, dann nennt man das „Process Operation“. Das bedeutet, dass immer wieder ein Prozess des gleichen Typs (z.B. Wareneingang) gelebt wird, und dass Sonderfälle (z.B. kaputte Ware) auch behandelt werden, selbst wenn dafür keine explizite Vorschrift existiert. Es ist die Prozessrealität. Der Kreis schließt sich, wenn dieser Prozess wieder analysiert wird um zu erfahren wie jetzt der Prozess gelebt wird, denn auf dem Weg durch die einzelnen Phasen kann sich sehr viel ändern. Der gelebte Prozess muss nicht mehr dem einmal entworfenen Prozess entsprechen. Geänderte Bedingungen, z.b. neue Gesetze oder neue Zulieferer, führen immer wieder zu Anpassungen des gelebten Prozesses. Um kontinuierlich die Prozesse eines Unternehmens zu überwachen und zu verbessern muss der Lebenszyklus immer wieder durchlaufen werden.

Fig. 7. BPM introduction used to condition participants, part 1/2

Das Prozessmodell als zentrales Arbeitsmittel
Ein Prozess ist die Gesamtheit aus Abläufen, Dokumenten, Entscheidungen, Personen und all den verbundenen Problemen. Um den Prozess zu visualisieren, wird Prozessmodellierung eingesetzt. Diese ist eine graphische Darstellung der Zusammenhänge. Im folgenden Bild2 sind Beispiele für die graphische Darstellung von Informationen bei der Prozessmodellierung aufgeführt.

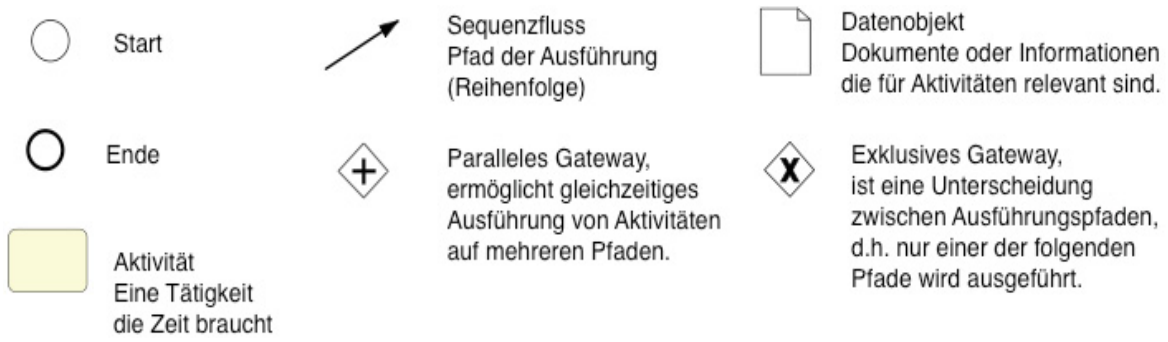


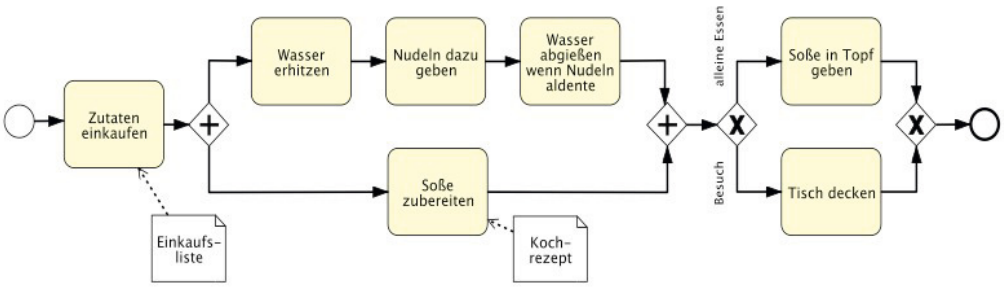
Bild2: Aspekte von Prozessmodellierung

Idealerweise werden die Prozessmodelle in allen Phasen als Kommunikationsmittel eingesetzt. Modelle können außerdem genutzt werden um Prozesse zu simulieren und so neue Abläufe vorher zu testen. Die Modelle werden auch als Vorlage bei der Softwareentwicklung eingesetzt oder dienen als graphisches Konfigurationswerkzeug für Standardsoftware.

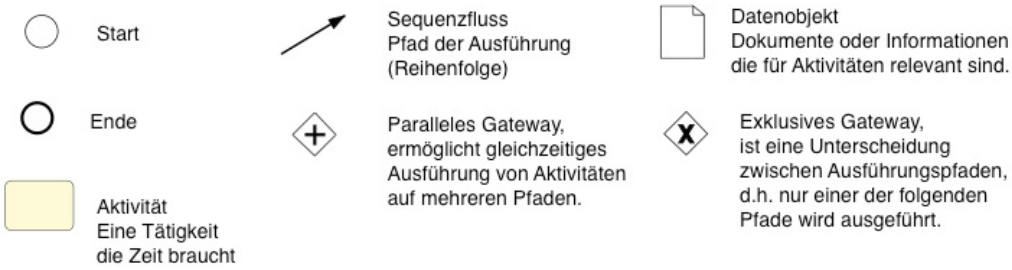
Die Arbeit des Lehrstuhls
Der Lehrstuhl für Business Process Technology lehrt und forscht in allen Bereichen des Prozesslebenszyklus. Im Besonderen geht es darum wie Prozessmodelle in Softwaresystemen eingesetzt werden können um Prozesse zu entwerfen, zu simulieren, umzusetzen und zu überwachen.

Fig. 8. BPM introduction used to condition participants, part 2/2

Beispiel für ein Prozessmodell



Legende der Symbole



Optimale Modellierung

- In Modellen werden oft nur die häufigsten Fälle abgebildet. Als Daumenregel gilt, dass 80% der realen Prozesse mit dem Model erfasst werden sollten.
- Zur besseren Lesbarkeit werden Aktivitäten mit *Objekt Verb* benannt werden, z.b. *Tisch decken* oder *Wasser erhitzen*
- Ausbalancierte Nutzung von Gateways: Wird ein Gateway genutzt, um Pfade aufzuteilen (alternativ/parallel), dann sollten die Pfade auch wieder mit einem Gateway vereint werden. Analog zum Klammern setzen in der Mathematik (siehe auch Beispiel oben).
- Entscheidungskriterien werden an die ausgehenden Pfade des exklusiven Gateways notiert (siehe Beispiel: *alleine Essen* vs. *Besuch*)

Fig. 9. BPMN Sample sheet used to introduce participants to process models

Kauf eines neuen Großbildschirms für den Eingangsbereich

Die Aufgabe
Der Unternehmenschef möchte Werbefilme im Foyer des Unternehmens laufen lassen. Dazu soll ein großer Bildschirm (mindestens 80") gekauft werden. Sie sind für die Beschaffung und Abrechnung des Bildschirms zuständig. Sie sind nicht für den Werbefilm oder die Installation im Foyer zuständig. Schildern Sie die Schritte die notwendig sind.
Beginnen Sie mit dem Moment indem Ihr Chef Ihnen die neue Aufgabe übertragen hat. Enden Sie, wenn alle Rechnungen bezahlt und abgeheftet sind. Wenn die Aufgabe Spielraum lässt, dann treffen sie sinnvolle Annahmen.

Ausschreibung eines neuen Lagergebäudes

Die Aufgabe
Die Firma expandiert. Ihr Chef möchte ein neues Lagergebäude für Reifen auf dem Werksgelände errichten lassen. Es sollen Angebote verschiedener Baufirmen einholt und verglichen werden. Sie sind für die Ausschreibung und die Begleitung des Projektes zuständig.
Beginnen Sie in dem Moment indem Ihr Chef Ihnen die neue Aufgabe übertragen hat. Ihre Beteiligung endet, wenn das Gebäude eingeweiht ist.

Fig. 10. Two samples for the introduction to the experimental task

Interview Guide:

Students get asked the following questions in exactly this wording and order by the experimenter:

- Frage1: Versuche alle relevanten Schritte zu identifizieren
- Frage2: Welche Dokumente spielen eine Rolle?
- Frage3: Gibt es grobe Phasen in deinem Vorgehen, die du identifizieren kannst?
- Frage4: Gibt es Schritte die nicht von einander abhängen so dass die Reihenfolge der Ausführung eigentlich egal ist, sie könnten also parallel ausgeführt werden?
- Frage5: Welche Probleme erwartest du bei diesem Prozess?
- Frage6: Gibt es noch etwas, dass du uns über diesen Prozess mitteilen möchtest?

When asked a question from the subjects. The experimenter shall use one of the following answers (if applicable):

- „Triff eine Annahme und gehe von dort weiter.“
- "Dazu gibt es vielleicht Hinweise in der Aufgabenstellung"
- "Dazu gibt es vielleicht Modellierungshinweise"
- "Das weiß ich nicht."

Fig. 11. Interview guide used by the experimenter to run the experimental tasks

		stimme gar nicht zu	stimme eher nicht zu	teils / teils	stimme eher zu	stimme voll zu
1	Ich konnte in der Feedback-Phase wichtige Anmerkungen machen.	□	□	□	□	□
2	Ich konnte mich für diese Methode der Prozesserhebung begeistern.	□	□	□	□	□
3	Diese Methode hat nichts zu meinem Wissen über Prozesse beigetragen.	□	□	□	□	□
4	Ich war motiviert, die Aufgabe zu erfüllen.	□	□	□	□	□
5	Ich bin mit meiner Lösung unzufrieden.	□	□	□	□	□
6	Mir war klar, was von mir erwartet wird.	□	□	□	□	□
7	Ich würde nicht noch einmal am Experiment teilnehmen.	□	□	□	□	□
8	Ich war mir bewusst, auf was die Aufgabe hinausläuft.	□	□	□	□	□
9	Ich habe durch diese Methode etwas über Prozesse dazugelernt.	□	□	□	□	□
10	Mir war es wichtig, in der Feedback-Phase mein Wissen über den Prozess einfließen zu lassen.	□	□	□	□	□
11	Es war mir egal, ob ich die Aufgabe gut löse oder nicht.	□	□	□	□	□
12	Ich konnte durch diese Methode mein generelles Prozessverständnis verbessern.	□	□	□	□	□
13	Ich bin von meiner Lösung überzeugt.	□	□	□	□	□
14	Diese Methode hat mir Spaß gemacht.	□	□	□	□	□
15	Mir war nicht immer klar, was genau ich tun soll.	□	□	□	□	□
16	Die Feedback-Phase empfand ich als unnötig.	□	□	□	□	□
17	Meine Lösung ist korrekt.	□	□	□	□	□
18	Diese Methode empfand ich als nervig.	□	□	□	□	□

Fig. 12. Questionnaire Document used. Three statements together operationalize one hypothesis.

Fig. 13. Feedback test: "Finding a new flat"

Fig. 14. Feedback test: "Getting a new job"

Aktuelle Technische Berichte des Hasso-Plattner-Instituts

Band	ISBN	Titel	Autoren / Redaktion
40	978-3-86956-106-6	**Selected Papers of the International Workshop on Smalltalk Technologies (IWST'10)**	Hrsg. von Michael Haupt, Robert Hirschfeld
39	978-3-86956-092-2	**Dritter Deutscher IPv6 Gipfel 2010**	Hrsg. von Christoph Meinel und Harald Sack
38	978-3-86956-081-6	**Extracting Structured Information from Wikipedia Articles to Populate Infoboxes**	Dustin Lange, Christoph Böhm, Felix Naumann
37	978-3-86956-078-6	**Toward Bridging the Gap Between Formal Semantics and Implementation of Triple Graph Grammars**	Holger Giese, Stephan Hildebrandt, Leen Lambers
36	978-3-86956-065-6	**Pattern Matching for an Object-oriented and Dynamically Typed Programming Language**	Felix Geller, Robert Hirschfeld, Gilad Bracha
35	978-3-86956-054-0	**Business Process Model Abstraction : Theory and Practice**	Sergey Smirnov, Hajo A. Reijers, Thijs Nugteren, Mathias Weske
34	978-3-86956-048-9	**Efficient and exact computation of inclusion dependencies for data integration**	Jana Bauckmann, Ulf Leser, Felix Naumann
33	978-3-86956-043-4	**Proceedings of the 9th Workshop on Aspects, Components, and Patterns for Infrastructure Software (ACP4IS '10)**	Hrsg. von Bram Adams, Michael Haupt, Daniel Lohmann
32	978-3-86956-037-3	**STG Decomposition: Internal Communication for SI Implementability**	Dominic Wist, Mark Schaefer, Walter Vogler, Ralf Wollowski
31	978-3-86956-036-6	**Proceedings of the 4th Ph.D. Retreat of the HPI Research School on Service-oriented Systems Engineering**	Hrsg. von den Professoren des HPI
30	978-3-86956-009-0	**Action Patterns in Business Process Models**	Sergey Smirnov, Matthias Weidlich, Jan Mendling, Mathias Weske
29	978-3-940793-91-1	**Correct Dynamic Service-Oriented Architectures: Modeling and Compositional Verification with Dynamic Collaborations**	Basil Becker, Holger Giese, Stefan Neumann
28	978-3-940793-84-3	**Efficient Model Synchronization of Large-Scale Models**	Holger Giese, Stephan Hildebrandt
27	978-3-940793-81-2	**Proceedings of the 3rd Ph.D. Retreat of the HPI Research School on Service-oriented Systems Engineering**	Hrsg. von den Professoren des HPI
26	978-3-940793-65-2	**The Triconnected Abstraction of Process Models**	Artem Polyvyanyy, Sergey Smirnov, Mathias Weske
25	978-3-940793-46-1	**Space and Time Scalability of Duplicate Detection in Graph Data**	Melanie Herschel, Felix Naumann
24	978-3-940793-45-4	**Erster Deutscher IPv6 Gipfel**	Christoph Meinel, Harald Sack, Justus Bross

ISBN 978-3-86956-108-0
ISSN 1613-5652