Tangible Media in 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 typically not involved in actual modeling. We created a tangible toolkit for process modeling to be used with domain experts. We hypothesize that it results in more effective process elicitation.

This paper assesses nine aspects related to "effective elicitation" in a controlled experiment using questionnaires and video analysis. We compare our approach to structured interviews in a repeated measurement design. Subjects were 17 student clerks from a trade school.

We conclude that tangible modeling leads to more effective elicitation through activation of participants and validation of results. In particular, subjects take more time to think about their process and apply more corrections to it. They also report to get insights into process modeling.

Keywords: process elicitation, tangible media, controlled experiment.

1 Introduction

In business process management graphically depicted process models serve as communication vehicles about the working procedures of organizations. They are the basis for a shared understanding and process improvements. Moreover, process models are often used as requirements engineering artifacts for software implementation projects. Supporting processes with software offers great potential to save time, enhance reliability and deliver standardized output [9]. At the same time, misunderstandings in early stages lead to expensive change requests at later stages of the software project. Thus, the quality of communication between stakeholders is crucial to translate process requirements into software implementation.

In current practice, process models are created by trained method experts, typically external consultants. They gather the required information in interviews or workshops with the stakeholders of the process [1]. Afterwards, the method expert creates a business process model using notations such as EPC or BPMN. Creation of process models is done with dedicated software.

Domain experts provide information upfront but are typically passive while their knowledge is translated into a process model by the modeling expert. This translation step undertaken by the modeling expert de-couples the domain expert

H. Mouratidis and C. Rolland (Eds.): CAiSE 2011, LNCS 6741, pp. 283–298, 2011.

from the model. When asked for feedback, additional effort is needed to explain the models meaning and to resolve misunderstandings. This paper addresses this problem by introducing an approach to couple domain experts with process models using tangible objects.

We have developed the tangible business process modeling (t.BPM) toolkit. 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 Model and Notation (BPMN), see Fig. 1. It consists of shapes for activities, gateways, events and data objects. Control flow and roles are drawn on the table. In our opinion, it enables domain experts to actively shape their processes and allows the method expert to act as a facilitator rather than a translator. 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.

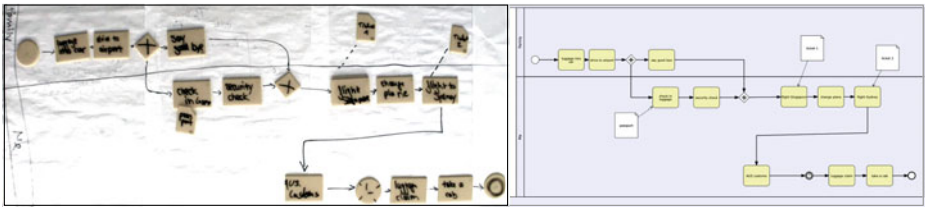

Fig. 1. Same process: modeled with t.BPM (left) and in a software modeling tool (right)

This paper reports on a controlled experiment in which we analyze one-to-one interview situations with respect of the effectiveness of process elicitation with or without t.BPM. It is a condensed version of a extensive technical report [15] on this experiment. Two hypotheses were cut out and discussed in detail in a separate publication [8]. Three more hypotheses were dropped for this paper as they did not hold and don't add value to the discussion here.

We review related research on process modeling in the next section. Afterwards, we explain the hypotheses, the experiment setup, the variables and the analysis procedures used in Section 3. The experiment execution is discussed in Section 4 and the data analysis is reported in Section 5. The results from the analysis are interpreted in Section 6 and the paper is concluded in Section 7.

2 Related Work

Empirical research on process modeling is typically focussed on the models that are produced with software tools and can be automatically analyzed, e.g. [5]. Only some research is turned towards the modelers in front of the screen and the process of model creation.

As examples, Sedera et al. [21] used case study research and survey methods to derive qualitatively a framework of factors that influence the success of process modeling efforts. Amongst others, they found tooling and participation to be key drivers. Participative model building was investigated by Persson [16]. She found that it leads to enhanced model quality, more stakeholder consensus and more commitment to results. The workshops are set up with a dedicated software tool operator to channel participant knowledge and create a common picture at the projector [23]. Rittgen developed software and guidance for modeling workshops in which the participants themselves use the software to create the model together [18]. Our approach uses intuitive tooling to remove software as a barrier for individuals to participate.

For individuals, Recker found that modeler performance is influenced by the complexity of the grammar [17], modeling experience and modeling background. Controlled experiments with individuals have been conducted e.g. by Weber et al. [24] to investigate the effect of events on process planing performance or by Holschke [10] to investigate the influence of model granularity on reusability of artifacts. To our best knowledge there is no controlled experiment that investigates the presence of an intuitive mapping tool for business process modeling.

The setup and execution of our controlled experiment was guided by Creswell [3] and Wohlin et al. [25]. We use literature from experimental software engineering [12] and statistics [6] to inform the structure of the paper and the level of reporting.

3 Experiment Planning

We outline all planning activities in this section. We start by deriving our hypotheses, talk about setup, the actual measurement of the hypotheses and the analysis procedures.

3.1 Goal and Hypotheses

The goal of this paper is to examine the effect of t.BPM on process elicitation with individuals. Therefore we compare t.BPM to structured interviews which are seen as the most effective requirements elicitation technique [4]. By 'effective' we mean that it produces a 'desired or intended result' [22]. In requirements engineering, more information is typically indicating more effective elicitation. But it was already shown that the presence of visual representations does not necessarily elicit more information [4]. We argue that effective process elicitation has more aspects such as user engagement and validated results. Fig. 2 visualizes how we refine our model towards hypotheses based on the following considerations:

User engagement is widely recognized as a key factor for success of IT projects [21]. Our approach uses tangible media which is seen as a key factor for task engagement, e.g. in HCI research [11]. In those cases, engagement is typically measured as the time spent on a problem, e.g. by Xie et al. [26]. Since tangible modeling consumes time to handle the tool itself (e.g, writing on tiles),

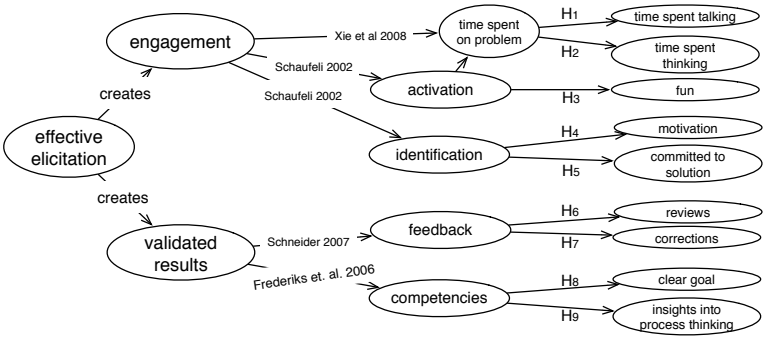

Fig. 2. Effective elicitation decomposed into nine hypotheses

we split up the time into more fine granular observations. We hypothesize that people will *spent more time talking (H_1)* about the process but also *spent more time to think (H_2)* about what they do.

Schaufeli segments engagement into the dimensions activation and identification [19]. While activation is already measured with the time spent on the task, we additionally hypothesize that people have *more fun (H_3)* as a further aspect of activation. The dimension of identification inspires us to hypothesize that people modeling with t.BPM will have *more motivation (H_4)* to accomplish the task and will be *more committed to the solution(H_5)* that they shaped.

The second key aspect that we see for effective elicitation is a validated result. Schneider [20] points out that validation cycles are a time consuming aspect of requirements elicitation projects. He proposes to create a model during the elicitation to trigger instant feedback and speedup validation. Validation cycles are characterized by reviews and adjustments to the model. We hypothesize that people will do *more reviews (H_6)* when using t.BPM and apply *more corrections (H_7)* to their process model.

Finally, validation in model building depends on the competencies of the participants. Frederiks [7] proposes that users validate models by deciding on the significance of information. We propose that this depends on a clear understanding of the modeling goal. We hypothesize that t.BPM provides a *clearer goal (H_8)* for the elicitation session. Frederiks also proposes that modeling experts guide the validation by grammatical analysis, in other words their modeling knowledge. We hypothesize that tangible modeling can create *insights into process thinking (H_9)* for the user and thereby support the validation process.

We note that the hypotheses are not a forced consequence of the identified aspects and their building involved interpretation. We come back to this decomposition when we assess the measurement validity in Section 5.3.

3.2 Experiment Setup and Sampling Strategy

We design the following experimental setup as illustrated in Fig. 3. Subjects get first conditioned to a certain level of BPM understanding. Afterwards they are

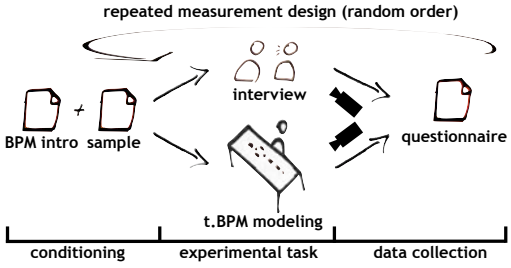

Fig. 3. Experiment Setup for this study

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. Two persons operate the experiment. One guides the subjects in the role of an interviewer, the other one 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, a questionnaire is to be filled in by the subjects. 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.

In other words, we use a randomized balanced single factor design with repeated measurements [25] also known as a within-subjects design. All subjects get both treatments assigned in different order. All subjects do interviews and process modeling. Finally, all subjects are rewarded for their participation with a chocolate bar and a cinema voucher.

3.3 Experimental Material

We briefly outline and explain the printed material used in this experiment. The original documents are appended to the technical report [15]. Like the experiment, the experimental material is in German.

- BPM introduction: Two pages explaining the terms Business Process Management, Business Process Modeling and process models.
- Sample model: One page depicting the process of "Making Pasta" and four pragmatical hints on process modeling such as verb-object style labeling.
- 2x task sheet: One paragraph explaining the process to report on. It explicitly sets the context, the start and the end-point of the process.
- Interview guide (for experimenter): Experimenters read out the same six questions in each experimental task. The sheet also contains standardized answers to potential questions from participants.
- Questionnaire: Items to be rated on a 5-point Likert scale. Details are explained in Section 3.5.

3.4 Participant Selection

The sample population used in research studies should be representatives of the population to which the researchers wish to generalize [2]. Thus, we want potential users of t.BPM to participate in the study. We got the opportunity to run an on-site experiment at a trade school in Potsdam (Germany) with graduate office and industrial clerk students. They all work in companies and study part time at the trade school. Industrial clerks do planing, execution and controlling of business activities. Office clerks do supporting activities in a department, e.g. as office managers. Both groups might be questioned in process-oriented projects by external consultants. Thus, they represent the target population that we like to address with t.BPM.

3.5 Operationalized Hypotheses

We operationalize the hypotheses presented in Section 3.1 by means of a questionnaire and video analysis. We define each hypothesis as H_x and its null hypothesis as H_{0x}.

Questionnaire-Based Hypotheses (H_3,H_4,H_5,H_8,H_9): Hypotheses which rely on perceived measures are tested using a questionnaire. On a five-point Likert scale, subjects rate their agreement to, in summary, fifteen statements. Three statements together represent one hypothesis. Two statements are formulated towards the hypotheses, one is negatively formulated. The level of agreement is mapped to the values [1..5] where 1 is no agreement and 5 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. We test the following hypotheses:

- H_3: Subjects report more fun in t.BPM sessions than in interviews.
  H_{03}: Subjects don't more fun in t.BPM sessions.
- H_4: Subjects report to be more motivated in t.BPM sessions than in interviews.
  H_{04}: Subjects don't report to be more motivated in t.BPM sessions.
- H_5: Subjects report to be more committed to the solution in t.BPM sessions than in interviews.
  H_{05}: Subjects report to be more committed to the solution in t.BPM sessions.
- H_8: Subjects report a clearer goal understanding in t.BPM sessions than in interviews.
  H_{08}: Subjects don't report a clearer goal understanding in t.BPM sessions.
- H_9: Subjects report to gain more new insights in process understanding from t.BPM sessions than from interviews.
  H_{09}: Subjects don't report to gain more new insights in process understanding from t.BPM sessions.

Video Hypotheses (H_1,H_2,H_6,H_7): We operationalize hypotheses related to time and actions taken during the experimental task using video coding analysis. We define the following coding schemes:

Time Slicing(H_1,H_2): The duration of the experimental task is sliced exclusively to belong to one of five categories. The use of t.BPM (Use_{tBPM}) 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 $Talk_{tBPM}$). We define a code for the time spent silent ($Silence_{tBPM/int}$) 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_6,H_7): Both are 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 repositioning that impacts the process model 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.

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

- H_1: Subjects talk more in t.BPM sessions than in interviews,
  i.e. $Talk_{tBPM} + UseTalk_{tBPM} > Talk_{int}$.
  H_{01}: Subjects don't talk more in t.BPM sessions,
  i.e. $Talk_{tBPM} + UseTalk_{tBPM} \not> Talk_{int}$.
- H_2: Subjects are more silent in t.BPM sessions than in interviews,
  i.e. $Silence_{t.BPM} > Silence_{int}$
  H_{02}: Subjects are not more silent in t.BPM sessions,
  i.e. $Silence_{t.BPM} \not> Silence_{int}$
- H_6: Subjects make more reviews in t.BPM sessions than in interviews,
  i.e. $Reviews_{t.BPM} > Reviews_{int}$
  H_{06}: Subjects don't make more reviews in t.BPM sessions,
  i.e. $Reviews_{t.BPM} \not> Reviews_{int}$
- H_7: Subjects make more corrections in t.BPM sessions than in interviews,
  i.e. $Corrections_{tBPM} > Corrections_{int}$.
  H_{07}: Subjects don't make more corrections in t.BPM sessions,
  i.e. $Corrections_{tBPM} \not> Corrections_{int}$.

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. We use a notational convention for the data sets collected: $intention_{V/Qx}$. As an example, $talking_{V1}$ describes the set of talking times as measured with the video analysis for hypothesis 1. Likewise, fun_{Q3} is the set of all ratings collected with the fun related questionnaire items for hypothesis 3.

3.7 Analysis Procedures

For hypothesis testing, we use a one-way repeated-measures ANOVA (analysis of variances). It aims to determine the variation within subjects that is caused by the method. Additionally, we carry out a dependent t-test with acceptance level $p<.05$. It is used to get a different view on the data and to assess potentially confounding factors that might have influenced the performance of the subjects.

To assess reliability of the questionnaire, we use Cronbach's alpha. It determines the internal consistency of the three questionnaire items measuring one hypothesis. The video data is analyzed by two independent reviewers. They compare their results and (if needed) resolve conflicts by negotiation. Cohen's Kappa is used to determine the inter-rater agreement before negotiation to assess the quality of our coding guidelines.

4 Experiment Execution and Data Collection

The experiment design was executed in December 2009 at a trade school in Potsdam. Slots were offered to the students by short teasers given in the classes. All subjects were at the age of nineteen to twenty-one. Students could choose to swap one lecture unit for experiment participation (about 1h). We expected to test industrial clerks only, but only ten volunteered. Thus, we opened up the experiment to office clerks as well. We ended up testing 7 office clerks and 10 industrial clerks within the week.

Fig. 4. Photos from the experiment execution. Subject giving interview (left) and modeling with t.BPM (right). Taped by the video cameras.

Each experiment run started with a short informal warm-up chat and afterwards 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. Fig. 4 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.

5 Data Analysis

We explain the analysis techniques used and the results found in this section. We reason about the results in Section 6.

5.1 Descriptive Statistics

From seventeen students in two runs, we collected 34 questionnaires with 510 statements in total. The video analysis is based on 6,74 hours of video material. One t.BPM session taping went wrong. That results in N=16 for all hypotheses that rely on 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.

5.2 Data Set Preparation

The data was tested with the Kolmogorov-Smirnov and Shapiro-Wilk test and is normally distributed. The original experiment evaluation involved two more video codings and three more questionnaire items. The related hypotheses did not hold and the data was therefore dropped for discussion in this paper due to the limited space. Apart from that, no collected data was excluded from the set.

5.3 Measurement Reliability and Validity

According to Kirk and Miller the reliability is the extent to which "a measurement procedure yields the same answer however and whenever carried out" ([13], p.19) while validity is the "extent to which it gives the correct answer".

We assess two aspects of measurement reliability. First we check the inter-rater agreement for the video codings using Cohen's kappa coeficient (κ). It compares both video codings before the negotiation process. The inter-rater agreement over all videos and all coding schemes is $\kappa = .463$ where $0.41 < \kappa < 0.60$ is a moderate agreement level [14]. Thus, we interpret our coding instructions as reasonably reliable and the results as moderately reproducible.

Furthermore, the reliability of the questionnaire is measured using Cronbach's alpha (α). It determines the degree to which the items related to one hypothesis coincide. In other words, whether they actually measure the same underlying concept, e.g. fun. In the literature [6] $\alpha > .8$ is suggested to be a good value for questionnaires, while $\alpha > 0.7$ is still acceptable. All our variables had $\alpha > .8$, except for $\alpha(motivation_{Q4}) = .702$ and $\alpha(clarity_{Q8}) = .687$. We keep those exceptions in mind but overall a high degree of reliability is indicated for the questionnaire.

Validating whether our variables correctly describe "effective elicitation" is not directly possible. We use effective elicitation as an umbrella term for the aspects of engagement and result validation. From there we derive variables to measure these aspects. In [15] we conducted a principal component analysis for validation. It is a technique to determine sets of strongly correlating variables which are approximated with one factor, the principal component [6]. Ideally, the variables would form two factors. Those that reflect the measures for engagement and those measuring result validation.

Using orthogonal (varimax) rotation, our nine dependent variables split up to three factors that do not match our hypothesis decomposition. Interestingly, all

questionnaire based variables aggregate to one large principal component. These measures rely on self-perception of the subjects and therefore describe one side of the coin. Moreover, the time for $talking_{V1}$ and $silence_{V2}$ strongly correlate with the amount of $reviews_{V6}$ done. It indicates the degree to which people were involved with the task. Finally, $corrections_{V7}$ builds a single factor. Overall, the measurement validation calls for a more thought-out hypothesis decomposition and clever selection of measurement instruments.

5.4 Hypothesis Testing

We use the repeated-measures ANOVA to determine the effect of our independent variable (method) within each individual per dependent variable. In other words, to what extend did the method influence the performance of each individual? Fig. 5 illustrates how our data is partitioned. 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).

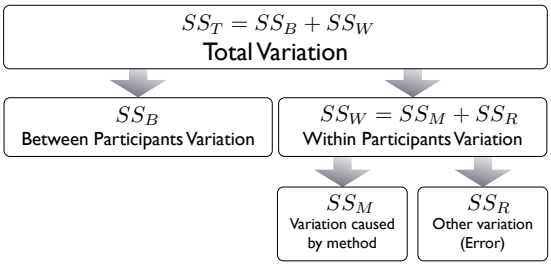

Fig. 5. Data partitioning for rep.-measures ANOVA. Drawing adopted from [6] p.463

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 $F_{.05}(1, 16) > 4.49$ 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 1 we sorted the variables according to descending $F_{.05}$ ratios. 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.

Furthermore we conduct a dependent t-test to create a different view on the data, see Table 2. It compares the groups doing t.BPM and interviews by their mean scores (V=in minutes, Q=Likert scale [1..5]), the statistical significance of this difference (one-tailed with acceptance level p<.05) and the confidence interval. The upper and lower boundaries indicate that the real mean difference

Table 1. ANOVA result table based on df_M=1. Sorted by $F_{.05}$ ratios

dependend Variable	df_R	SS_T	SS_B	SS_M	SS_R	$F_{.05}$	η^2
$corrections_{V7}$	15	119.22	42.72	57.78	18.72	**46.30**	0.76
$silence_{V2}$	15	398.55	129.58	167.92	101.05	**24.93**	0.62
$insights_{Q9}$	16	18.24	14.9	0.84	2.50	**5.36**	0.25
$reviews_{V6}$	15	38.01	23.00	3.13	11.88	3.95	0.21
$talking_{V1}$	15	116.56	56.92	10.86	48.79	3.34	0.18
fun_{Q3}	16	18.31	15.03	0.55	2.73	3.24	0.17
$commitment_{Q5}$	16	24.68	20.90	0.33	3.45	1.52	0.09
$clarity_{Q8}$	16	32.78	25.78	0.12	6.88	0.27	0.02
$motivation_{Q4}$	16	10.90	9.46	0.05	1.39	0.23	0.04

Table 2. (one-tailed) t-test comparing groups by method. Ordered like Table 1

dependent variable	Effect Size t.BPM	Effect Size interview	Significance	Confidence Intervals lower boundary	Confidence Intervals upper boundary
$corrections_{V7}$	3.00	0.31	**.000**	**1.85**	**3.53**
$silence_{V2}$	5.54	0.95	**.000**	**2.63**	**6.54**
$insights_{Q9}$	3.75	3.43	**.017**	**0.03**	**0.60**
$reviews_{V6}$	0.81	0.19	**.033**	-.046	1.30
$talking_{V1}$	4.65	3.49	**.044**	-0.19	2.52
fun_{Q3}	4.16	3.90	**.046**	-0.05	0.56
$commitment_{Q5}$	3.31	3.51	.118	-0.53	0.14
$clarity_{Q8}$	3.37	3.49	.304	-0.59	0.36
$motivation_{Q4}$	4.45	4.37	.225	-0.14	0.29

between the groups is in that range with 95 percent probability. It should not include zero to be sure about the effect between the groups.

From both tables we see, that all parameters for $corrections_{V7}$, $silence_{V2}$ and $insights_{Q9}$ meet scientific standards. For $reviews_{V6}$, $talking_{V1}$ and fun_{Q3} we see that they just missed acceptable standards in both tests. E.g. the difference between the groups is significant in Table 2 but the confidence intervals do not allow acceptance by rigor scientific standards. Finally, $commitment_{Q5}$, $clarity_{Q8}$ and $motivation_{Q4}$ did not hold.

5.5 Testing Potentially Influential Factors

We use a two-tailed dependent t-test to compare groups were two different influences were applied. For example, we had two processes to report on, two experimenters, and two different educational groups. Furthermore, each subject goes through the experimental task twice. Repetition effects might have influenced the performance of the subjects.

While the experimenters had no significant influence on the dependent variables, we found that the second experimental task led to significantly more $clarity_{Q8}$ about the goal (1st=3.1, 2nd=3.77, p=.001) and more $commitment_{Q5}$ to the solution (1st=3.2, 2nd=3.63, p=.004).

Participants' education had significant influence on $clarity_{Q8}$ and $insights_{Q9}$. In particular, office clerks reported to have a clearer goal understanding (o-clerks=3.98, i-clerks=3.05, p=.031) and get more new insights into process thinking (o-clerks=4.05, i-clerks=3.30, p=.022). In all cases, the confidence intervals left no doubt about the effect.

6 Interpretation of Results

We can identify three types of variables. Those that support their hypothesis, those that do not support their hypothesis, and those that just missed rigor scientific standards. We consider the latter ones as conditionally supportive and argue that a slightly larger sample set would have made the difference.

This claim is based on the t-test in Table 2. It indicates a significant difference for $talking_{V1}$,fun_{Q3}, and $reviews_{V6}$ due to method. The confidence intervals do not allow acceptance with scientific rigor. That means, we cannot rule out with 95 percent probability that the actual effect size is zero. For example, $talking_{V1}$ time is significantly higher (p=.044) in t.BPM sessions (t.BPM=4.65min, int=3.49min) but the confidence interval includes zero (lb=-0.19min, ub=2.52min). In this case we miss rigor acceptable levels by twelve seconds. The rest of the discussion is structured according to the hypothesis decomposition in Fig. 2.

The engagement variables to measure activation indicate a positive effect through method. Participants in t.BPM sessions did spend more time talking ($F_{0.5}(1, 15) = 3.34, \eta^2 = 0.18$) and significantly more time thinking ($F_{0.5}(1, 15) = 24.93, \eta^2 = 0.62$) about their process. They also report more fun ($F_{0.5}(1, 16) = 3.24, \eta^2 = 0.17$) in t.BPM sessions. We reject H_{02} and argue that H_{01} and H_{03} might be rejected with a bigger sample size.

The engagement variables measuring the dimension of identification did not hold. People did not report significantly more motivation or commitment to their solution due to the method used. We assume a ceiling effect for $motivation_{Q4}$. Participants got off from class, plus a chocolate bar and a cinema voucher for compensation. On a five point Likert scale we could not find a statistically relevant difference in $motivation_{Q4}$ due to the method applied (t.BPM=4.45, int=4.37). For $commitment_{Q5}$ we found in Section 5.5 that it significantly raises with repetition (p=.004). Thus, we assume that commitment (as operationalized by us) indicates self-confidence that raises with due to the learning effect. We do neither reject H_{04} nor H_{05}.

The variables that operationalize the aspect of validated results show a mixed picture. We note more reviews ($F_{0.5}(1, 15) = 3.95, \eta^2 = 0.21$) and significantly more corrections ($F_{0.5}(1, 15) = 46.3, \eta^2 = 0.76$) due to the method. We reject H_{07} and argue that H_{06} might be accepted with a sightly larger sample size. We conclude that t.BPM provokes more feedback in process elicitation sessions.

The competencies required for result validation rely on perceived measures. We see that people report significantly more insights into process thinking ($F_{0.5}(1, 15) = 5.36, \eta^2 = 0.25$) in t.BPM sessions but the goal clarity does not

raise likewise ($F_{0.5}(1, 15) = 0.27$, $\eta^2 = 0.02$). In Section 5.5 we found that goal clarity significantly raises with repetition (p=.001). We conclude that, similar to commitment, the goal clarity is determined by learning rather than the method. We reject H_{09} but not H_{08}.

In summary, we interpret the t.BPM method to be engaging through activation of subjects. We can not reason on the concept of identification which was determined by other effects in this experiment. The t.BPM method also leads to validated results through more feedback on the model. The competencies for result validation raise partially with the method and partially with learning through repetition.

6.1 Validity Threats

The **internal validity** was addressed by the experiment design. In particular, we use two processes and two experimenters assigned in random order. In Section 5.5 we assess potentially confounding variables for their influence. While experimenters and processes did not harm the results, we found learning effects due to the repeated measurements design on $clarity_{Q8}$ and $commitment_{Q5}$.

We found education to be influential on the reported $clarity_{Q8}$ and $insights_{Q9}$. In short, office clerks tend to report better scores while scoring worse in objective tasks [8]. While group heterogeneity is a threat to the internal validity, it also increases the **external validity** as both groups represent the population that we address with our tool. This is as important as choosing domain processes rather than artificial graphs to test with. We chose the domain processes in coordination with the school to ensure all students are equally familiar with them. However, we did not assess to which extend individuals are exposed to these processes in their companies. The **measurement instruments** were tested in one pre-study with ten computer science students. Small adjustments were made afterwards. To ensure quality standards for data analysis, we used two independent coders for the video analysis and we have split each questionnaire variable into three items, one poled negatively. Finally, we provide a longer version of this paper including more data and the experimental material in [15].

6.2 Generalizability of Findings

We think the findings about t.BPM can be generalized from the sample group to the general population. Besides their age (19-21years) the students represent exactly the group we address with the t.BPM tool.

We also think that the findings will hold for other tangible modeling approaches when compared to pure talking. Some aspects have also been reported for visual mappings of requirements such as instant feedback [20]. However, a different tests would be needed to determine exactly the aspects that lead to activation of participants.

6.3 Lessons Learned

If we had to start over again, we would put more effort into the reliability of our questionnaire items, in particular $clarity_{Q8}$. But we also learned that people may

report a glorified self-image. Thus, we suggest to mix measurement instruments for each measured concept. In other words, complement perceived measures with external measures such as video codings. But we also had to learn that rigor video analysis is the most time-consuming evaluation task.

Besides all that, we think that 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 to review each others work did ensure a stable setup.

7 Conclusion

This paper reports on a controlled experiment which was conducted with 17 student clerks at a trade school. We investigate the process elicitation method as an independent variable. Subjects did structured interviews and t.BPM in a repeated measurement design. We claim that t.BPM enables more efficient process elicitation. We argue that efficient elicitation is not about the amount of information but about user engagement and validated results. We decompose these aspects into nine operationalized hypotheses. Three hypotheses did hold. Three more might hold with a larger sample set.

The results show strong support for user engagement through activation of participants and validated results through more feedback from participants. We think that these findings are reproducible with other tangible system modeling approaches when compared with interviews.

Our findings are limited by the measurement instruments and the small sample size (N=17). A future experiment with a larger group and better tested instruments might re-enforce our findings and also support H_1,H_3 and H_6. In other words, it would extend our rigor findings to more talking, more fun and more reviews with tangible media. For now we only showed significantly more thinking time (H_2), more corrections (H_7) and more insights into modeling (H_9) when using tangible media instead of interviews.

Acknowledgements

We are grateful to the students that helped setting up, running, and evaluating this experiment, namely Karin Telschow, Markus Güntert and Carlotta Mayolo. We'd also like to thank the reviewers for their valuable feedback. It led to a substantial revision of Sections 3.1 and 6.

References

1. Byrd, T., Cossick, K., Zmud, R.: A synthesis of research on requirements analysis and knowledge acquisition techniques. MIS Quarterly, 117–138 (1992)
2. Cooper, D., Schindler, P.: Business Research Methods, 10th edn. McGraw-Hill Higher Education, New York (2008)

3. Creswell, J.W.: Research design: Qualitative, quantitative, and mixed methods approaches. Sage Pubns, Thousand Oaks (2008)
4. 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, pp. 179–188 (2006)
5. van Dongen, B., van der Aalst, W., Verbeek, H.: Verification of ePCs: Using reduction rules and petri nets. In: Pastor, Ó., Falcão e Cunha, J. (eds.) CAiSE 2005. LNCS, vol. 3520, pp. 372–386. Springer, Heidelberg (2005)
6. Field, A.: Discovering statistics using SPSS. SAGE publications Ltd, Thousand Oaks (2009)
7. Frederiks, P.J.M., Van der Weide, T.P.: Information modeling: the process and the required competencies of its participants. Data & Knowledge Engineering 58(1), 4–20 (2006)
8. Grosskopf, A., Weske, M.: On business process model reviews. In: CAiSE 2010, pp. 31–42. Springer, Heidelberg (2010)
9. Hammer, M., Champy, J.: Reengineering the corporation: A manifesto for business revolution. Collins Business (2003)
10. Holschke, O., Rake, J., Levina, O.: Granularity as a cognitive factor in the effectiveness of business process model reuse. In: Dayal, U., Eder, J., Koehler, J., Reijers, H.A. (eds.) BPM 2009. LNCS, vol. 5701, pp. 245–260. Springer, Heidelberg (2009)
11. Ishii, H., Ullmer, B.: Tangible bits: towards seamless interfaces between people, bits and atoms. In: SIGCHI, pp. 234–241. ACM, New York (1997)
12. Jedlitschka, A., Ciolkowski, M., Pfahl, D.: Reporting experiments in software engineering. Guide to Advanced Empirical Software Engineering, 201–228 (2008)
13. Kirk, J., Miller, M.: Reliability and validity in qualitative research. Sage Publications, Inc., Newbury Park (1986)
14. Landis, J., Koch, G.: The measurement of observer agreement for categorical data. Biometrics 33(1), 159–174 (1977)
15. Luebbe, A., Weske, M.: The effect of tangible media on individuals in business process modeling - a controlled experiment. Tech. Rep. 41, Hasso-Plattner-Institute for IT Systems Engineering (2010), `http://bpt.hpi.uni-potsdam.de/Public/AlexanderGrosskopf`
16. Persson, A.: Enterprise modelling in practice: situational factors and their influence on adopting a participative approach. Ph.D. thesis, Stockholm University (2001)
17. Recker, J., Rosemann, M.: The measurement of perceived ontological deficiencies of conceptual modeling grammars. Data & Knowledge Engineering (2010)
18. Rittgen, P.: Success factors of e-collaboration in business process modeling. In: Pernici, B. (ed.) CAiSE 2010. LNCS, vol. 6051, pp. 24–37. Springer, Heidelberg (2010)
19. 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), 71–92 (2002)
20. Schneider, K.: Generating Fast Feedback in Requirements Elicitation. In: Sawyer, P., Heymans, P. (eds.) REFSQ 2007. LNCS, vol. 4542, pp. 160–174. Springer, Heidelberg (2007)
21. Sedera, W., Gable, G., Rosemann, M., Smyth, R.: A success model for business process modeling: findings from a multiple case study. In: PACIS, Shanghai (2004)
22. Stevenson, A.: Oxford Dictionary of English, vol. 24. Oxford University Press, Oxford (2010)

23. Stirna, J., Persson, A., Sandkuhl, K.: Participative Enterprise Modeling: Experiences and Recommendations. In: Krogstie, J., Opdahl, A.L., Sindre, G. (eds.) CAiSE 2007 and WES 2007. LNCS, vol. 4495, pp. 546–560. Springer, Heidelberg (2007)
24. Weber, B., Pinggera, J., Zugal, S., Wild, W.: Handling events during business process execution: An empirical test. In: ER-POIS at CAISE, pp. 19–30 (2010)
25. Wohlin, C., Runeson, P., Höst, M.: Experimentation in software engineering: an introduction. Springer, Netherlands (2000)
26. 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: Proceedings of TEI, pp. 191–198. ACM, New York (2008)