Towards Validating Prediction Systems for Process Understandability: Measuring Process Understandability (Experimental Results)

Joachim Melcher and Detlef Seese
Institut AIFB, Universität Karlsruhe (TH)
76128 Karlsruhe, Germany
{melcher|seese}@aifb.uni-karlsruhe.de

August 27, 2008

Abstract

Motivated by software metrics, several process metrics for measuring internal (structural) process attributes have been proposed. Integrated in validated *prediction systems, these metrics can be used to predict values of external process attributes like, for example, process understandability.*

There are only a few papers dealing with finding a process understandability metric and validating prediction systems for process understandability. Looking at these publications, we identified possible problems with metric reliability and validity.

In this paper, we define new metrics for process understandability inspired by existing work. Furthermore, we present some hypotheses about effects of measuring process understandability. Conducting an experiment, we got some encouraging findings supporting these hypotheses: Different aspects of process understandability can be complicated in varying degrees for a process and asking only some few questions about a process can cause values for process understandability differing very much from the real value. These findings should be considered in future work about measuring process understandability.

1 Introduction

Motivated by the research in software metrics, several papers proposing numerous process metrics for measuring internal (often structural) process attributes have been proposed (see, e. g., [6] for on overview). Integrated in *validated* prediction systems, these process metrics can be used to predict the values of external process attributes (e. g., error-proneness, time, costs) even before a process has been implemented or for external attributes that are measurable only in a very laborious manner.

One very important external attribute is process understandability by involved humans (e. g., process designers, process analysts, process implementers or people executing a process). Understandability influences other quality aspects of processes like error-proneness and maintainability. Even though the importance of understandability is undoubted, Mendling *et al.* state that "we know surprisingly little about the act of modeling and which factors contribute to a 'good' process model in terms of human understandability" [7, p. 48].

For examining process understandability and validating appropriate prediction systems, we first have to quantify process understandability. So, we have to find a proper process understandability metric fulfilling the reliability and validity requirements for metrics. Looking at the few existing works about the validation of prediction systems for process understandability, doubts especially about the reliability and validity of the used understandability metrics and the experimental design arise.

In this article, we give concrete and detailed definitions for measuring process understandability exceeding those in existing publications. Using these definitions, we formulate hypotheses about effects of measuring process understandability that have to be considered in the measuring process. An experimental evaluation

is conducted to examine these hypotheses.

The remainder of this paper is organized as follows: In Section 2, we present related work on measuring process understandability and validating prediction systems for understandability. Important basics about measurement and prediction systems are shown in Section 3. After giving comments and criticism on the existing approaches in Section 4, we present a framework for evaluating modeling technique understanding (Section 5). Our definitions for measuring process understandability and hypotheses about some effects of measurement are shown in Section 6. In Section 7, an experiment for examining these hypotheses is presented. The paper gives a conclusion and presents possible future work (Section 8).

2 Related Work

In [7], Mendling *et al.* searched for possible relations between personal and process specific (structural) properties and process understandability.

They used a questionnaire which was answered by 73 students having followed courses on process modeling. For the questionnaire, they selected 12 processes (each with 25 tasks). The processes were depicted in a simplified EPC-like notation (without events) in a top-to-bottom-style. The tasks were labeled with just capital letters.

Every student evaluated the perceived difficulty of each of the 12 processes (metric PERCEIVED). As operationalization of process understandability, Mendling *et al.* created the SCORE metric: Each student had to answer eight closed questions about order, concurrency, exclusiveness or repetition of tasks as well as one open question about possible errors for each process. The sum of correct answers (at most nine) gives the SCORE value.

Mendling *et al.* obtained the following results: There is only a loose relation between PERCEIVED and SCORE. Out of 20 process metrics, only DENSITY and AVERAGE CONNECTOR DEGREE have statistically significant correlations with SCORE. Furthermore, they created a linear regression model.

In [11], Vanderfeesten *et al.* introduced the cross-connectivity metric (CC). It was added as 21st process metric into the data collected in [7]. The correlation between CC and SCORE is *not* significant. Yet, a better regression model including the CC metric has been found.

In addition to the goals of [7], Mendling and Strembeck examined also the influence of content related factors on process understandability in [8].

For that purpose, they designed an online questionnaire that was answered by 42 students and practitioners. Six processes with equal number of tasks—each in two variants (one with tasks labeled with capital letters and a second one with tasks labeled with normal describing text)—were selected. The processes were depicted in the same notation as in [7]. For each process, six yes/no questions about process structure and behavior were chosen. The subjects of the experiment were randomly assigned to one of two questionnaire variants (capital letter labels and text labels).

The metric PSCORE was calculated as the sum of correct answers about the processes (at most 36) and served as an operationalization of understandability to a person.

For each process, the values of eight process metrics (SIZE, DIAMETER, STRUCTUREDNESS, SEPARABILITY, CYCLICITY, HETEROGENEITY and SOUND) were computed. The metric MSCORE was calculated as the sum of correct answers from all participants to one process. It served as an operationalization of understandability related to a process. The metric TEXTLENGTH was used to measure the string length of all textual task labels.

Mendling and Strembeck obtained the following results: Only SEPARABILITY has a significant Spearman correlation with MSCORE. The sum of correct answers does not significantly differ between the different label variants. Yet, there is a strong negative Pearson correlation between TEXTLENGTH and the number of correct answers.

3 Measurement and Prediction Systems

3.1 Definitions

The area of process measurement is inspired by the works and results of software measurement. There, many theoretical fundamentals were identified as important. Fenton and Pfleeger give a good overview in [2]. In [6], we show that these theoretical basics are also essential for process measurement—even so we had to notice that many of these findings are still ignored.

According to Fenton and Pfleeger, there are two main types of measurement:

Definition 1 (Measurement systems) *Measurement systems are used to assess an* existing *entity by numerically characterizing one or more of its attributes [2, p. 104].*

Definition 2 (Prediction systems) *Prediction systems are used to predict some attribute of a* future *entity, involving a mathematical model with associated prediction procedures [2, p. 104].*

Besides the use for *future* entities as stated in the definition of Fenton and Pfleeger, prediction systems can

also be used to predict some attribute of an *existing* entity that is measurable only in a very laborious manner.

In [6], we show how the idea of prediction systems can be transfered to process measurement (see Figure 1):

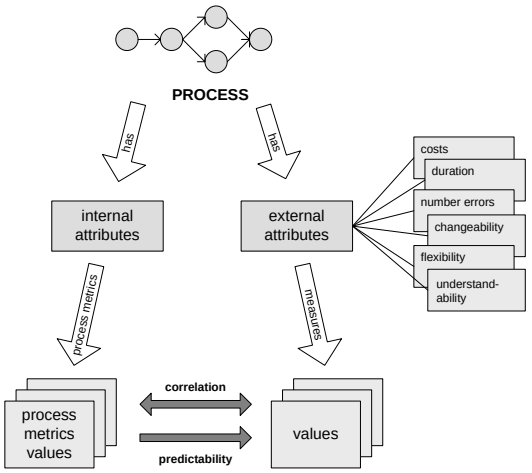


Figure 1: Prediction systems adapted to process measurement.

A process has *internal* and *external* attributes.

Internal attributes are those that can be measured purely in terms of the process separate from its behavior [2, p. 74]. Most proposed process metrics measure structural properties (internal attributes).

External attributes are those that can be measured only with respect to how the process relates to its environment [2, p. 74]. Examples are costs, time, number of errors and—especially important for this paper—understandability.

3.2 Validation

Before a prediction system can be used, it has to be validated. A valid prediction system consists of two metrics both being valid measurement systems. Valid measurement systems must fulfill the following two properties:

Reliability Metric values obtained by different observers of the same process have to be consistent. Kan gives a good example [4, pp. 70–71]: If one wants to measure the height of a person, the measurements should be taken at a special time of day (e. g., in the morning) and always barefooted. Otherwise, the values of the same person could vary a lot.

Validity According to Kan [4, pp. 71–72], validity can be classified into *construct validity* and *content validity*. The first checks whether the metric really represents the theoretical concept to be measured (e. g., is church attendance a good metric for religiousness?). The second checks whether the metric covers the range of meanings included in the concept (e. g., a test of mathematical ability for elementary pupils cannot be limited to addition but should also include subtraction, multiplication, division and so forth).

The goal of a validation of a prediction system is to show a correlation between the process metric values and the corresponding external attribute in question. As Fenton and Pfleeger state, "rather than being a mathematical proof, validation involves confirming or refuting a hypothesis" [2, p. 104].

4 Comments and Criticism on Existing Work

The observed loose relation between perceived and "objectively" measured[1] understandability [7, pp. 55–56] points out the need for an *operationalization* of process understandability. But the published work [7, 8] still gives reason for criticism:

Many details about the conducted experiments are not given in the papers (e. g., used processes, asked questions and measured metric values). This information is essential for replications of the experiments (by other researches) which is an important scientific methodology.

In the used experimental designs, different influencing factors (e. g., process metrics) are changed simultaneously and non-systematically. This could cause interactions between the different factors that influence understandability. So, the impact of one single changing process metric cannot be analyzed. The standard approach for controlled experiments uses *factorial design* (see, e. g., [2, pp. 135–146]). Here, all independent variables (process metrics in our case) are changed for their own while all others are kept constant.

In many places, only Pearson's correlation coefficients (measuring *linear* correlation) and generally only linear regression models are used to analyze possible relations between process metrics and process understandability. But there are many other imaginable (non-linear) kinds of correlation that will not be found this way! One possible solution to this problem is to use scatter plots to visually search for any kind of relation.

From our point of view, the biggest point to criticize is the possible lack of validity and reliability (see Subsection 3.2) of the metrics SCORE and MSCORE used to measure understandability.

[1] Although the SCORE and MSCORE metric values in [7, 8] are derived from answers given by humans (subjective influence), they are much more objectively than just asking for perceived understandability: Answers to concrete questions about a process have to be given and the proportion of correct answers can be determined objectively.

Content validity is about whether the metric covers the range of meanings included in the underlying concept. Mendling *et al.* name four aspects of process understandability: understanding of order, concurrency, exclusiveness and repetition [7, p. 52]. In [8, p. 146], Mendling and Strembeck ask questions about choices, concurrency, loops and deadlocks. But these are not used to compute the MSCORE metric for processes.

Looking at these publications, some questions arise: Do other important aspects of process understandability exist? How different is the understanding based on the different aspects? How can "overall process understandability" be computed?

Reliability requires that metric values obtained by different observers of the same process have to be consistent. In [7, 8], only eight and six questions per process are asked, respectively. And in [7], these questions are even distributed to four different aspects. It does not become clear how the process tasks involved in the questions are selected. So, the question arises if this selection is representative for the process. Maybe, complicated parts of the process have been omitted or only especially complex parts have been selected. In our opinion, this selection has a big influence on the measured values.

5 Framework for Evaluating Modeling Technique Understanding

Based on work by Mayer [5], Gemino and Wand proposed a framework for evaluating model understanding for arbitrary modeling techniques [3].

They differentiate between model creation (for representing parts of the real world) and model reading (creating a mental representation from a model) [3, p. 80]. In this paper, we deal with the second point.

For this purpose, they suggest a model for knowledge construction and learning from models adapted from Mayer: Content, presentation method and the model viewer characteristics influence the knowledge construction and consequently the learning outcome. This cognitive process is not directly observable, but has to be observed indirectly through learning performance tasks. Here, Gemino and Wand list comprehension and problem-solving tasks. The former include questions regarding attributes of and relationships between model items—while the latter include questions going beyond the information given originally in the model. [3, pp. 82–83]

For our problem (process understandability), comprehension tasks seem to be obvious.

6 Measuring Process Understandability

6.1 Aspects of Process Understandability

As we already discussed in Section 4, it is important to cover the different aspects of process understandability to fulfill the content validity requirement for metrics. In this paper, we concentrate on the aspects *order*, *concurrency*, *exclusiveness* and *repetition* identified by Mendling *et al.* in [7, p. 52]. Doing so, we do not deny the possible existence of other aspects. Unlike in [7], we will give detailed definitions of the questions of the different aspects.

We start with the definition of the term "activity period" which is later used in our questions.

Definition 3 (Activity period) *An activity period of task t is the period between a point in time when t becomes executable and the next point in time when the actual execution of t terminates.*

Now, we can define relations for the four aspects of process understandability.

Definition 4 (Order) *For the questions about task order, the relations $o_{\nexists}, o_{\exists}, o_{\forall} \subseteq T \times T$ with the following meanings are used.*

$(t_1, t_2) \in o_{\nexists} \Leftrightarrow$ There is no *process instance for which an activity period of task t_1 ends before an activity period of task t_2 starts.*

$(t_1, t_2) \in o_{\exists} \Leftrightarrow$ There is a *process instance for which an activity period of task t_1 ends before an activity period of task t_2 starts.—But there also exists a process instance for which this does not hold.*

$(t_1, t_2) \in o_{\forall} \Leftrightarrow$ For each *process instance, an activity period of task t_1 ends before an activity period of task t_2 starts.*

Definition 5 (Concurrency) *For the questions about task concurrency, the relations $c_{\nexists}, c_{\exists}, c_{\forall} \subseteq T \times T$ with the following meanings are used.*

$(t_1, t_2) \in c_{\nexists} \Leftrightarrow$ There is no *process instance for which the activity periods of tasks t_1 and t_2 overlap.*

$(t_1, t_2) \in c_{\exists} \Leftrightarrow$ There is a *process instance for which the activity periods of tasks t_1 and t_2 overlap at least once (Several executions of t_1 and t_2 per process instance are possible!).—But there also exists a process instance for which this does not hold.*

$(t_1, t_2) \in c_{\forall} \Leftrightarrow$ For each *process instance, the activity periods of tasks t_1 and t_2 overlap at least once.*

Definition 6 (Exclusiveness) *For the questions about task exclusiveness, the relations $e_{\nexists}, e_{\exists}, e_{\forall} \subseteq T \times T$ with the following meanings are used.*

$(t_1, t_2) \in e_{\nexists} \Leftrightarrow$ There is no *process instance, for which tasks t_1 and t_2 are both executed.*

$(t_1, t_2) \in e_\exists \Leftrightarrow$ *There is* a *process instance, for which tasks t_1 and t_2 are both executed.—But there also exists a process instance for which this does not hold.*

$(t_1, t_2) \in e_\forall \Leftrightarrow$ *For* each *process instance, the tasks t_1 and t_2 are both executed.*

Definition 7 (Repetition) *For the questions about task repetition, the relations $r_{=1}, r_?, r_*, r_+ \subseteq T$ with the following meanings are used.*

$t \in r_{=1} \Leftrightarrow$ *For* each *process instance, task t is executed* exactly once.

$t \in r_? \Leftrightarrow$ *For* each *process instance, task t is executed* not once or exactly once. *Both cases really occur.*

$t \in r_* \Leftrightarrow$ *For* each *process instance, task t is executed* not once, exactly once or more than once. *There exists a process instance for which t is executed not once and another one for which t is executed more than once.*

$t \in r_+ \Leftrightarrow$ *For* each *process instance, task t is executed* at least once. *There exists a process instance for which t is executed more than once.*

The definitions of the relations might look a little complicated. But we constructed them in such a way that we get the properties of Corollary 1, which are beneficial for the measurement process.

Corollary 1 (Properties of relations) *The relations have the following properties:*

1. *The relations $c_\sharp, c_\exists, c_\forall$ and $e_\sharp, e_\exists, e_\forall$ are symmetric.*
2. *For all possible task combinations, exactly one relation per aspect is true.*

Because of property 2 of Corollary 1, we can group the different relations for an aspect to questions about the process: The question $q_r(t)$, for example, asks which of the relations $r_{=1}, r_?, r_*, r_+$ holds for task t. Because of property 1 of Corollary 1, $q_c(t_1, t_2) = q_c(t_2, t_1)$ and $q_e(t_1, t_2) = q_e(t_2, t_1)$ hold.

Corollary 2 (Maximum number of questions) *The maximum number $|Q_{a,max}(p)|$ of possible different questions of aspect $a \in \{o, c, e, r\}$ about a process p with n tasks is*

$$|Q_{o,max}(p)| = n(n-1) \quad (1)$$

$$|Q_{c,max}(p)| = |Q_{e,max}(p)| = \frac{n(n-1)}{2} \quad (2)$$

$$|Q_{r,max}(p)| = n \quad . \quad (3)$$

As one can see, the maximum number of questions for *order*, *concurrency* and *exclusiveness* grows quadratically with the number of tasks, while the maximum number of questions for *repetition* grows only linearly.

We can now define process understandability.

Definition 8 (Personal process understandability) *The personal process understandability $U_a(p, s)$ of aspect a of process p by subject s is defined as the fraction of correct answers given by s to the $|Q_{a,max}(p)|$ different questions of aspect a about p.*

$$U_a(p, s) := \frac{\text{\# correct answers to } Q_{a,max}(p)}{|Q_{a,max}(p)|} \quad , a \in \{o, c, e, r\} \quad (4)$$

Hypothesis 1 *The personal process understandability metric values $U_a(p, s_i)$ of a process p are normally distributed.*

The different values of personal process understandability can be seen as outcomes of a random variable. The expected value of this variable can be estimated according to Definition 9.

Definition 9 (Estimated process understandability) *The estimated process understandability $\widehat{U}_a(p, S)$ of aspect a of process p and subjects S is defined as the average personal process understandability of p by the subjects of S.*

$$\widehat{U}_a(p, S) := \frac{1}{|S|} \sum_{s \in S} U_a(p, s) \quad , a \in \{o, c, e, r\} \quad (5)$$

Additionally, confidence intervals for the true expected values of the random variables for the different aspects of process understandability can be computed. The width of these intervals will decrease for higher numbers of subjects—meanwhile, the certainty of the true expected value will increase.

Hypothesis 2 *The different aspects of process understandability result in different values of the $\widehat{U}_a(p, S)$ of a process p.*

Consequently, it is important to measure all aspects to get "overall understandability". Furthermore, it should be examined whether other aspects of process understandability exist and how "overall understandability" can be computed from the values of the different aspects.

6.2 Partial Process Understandability

In order to reduce the effort for measuring process understandability, only a subset of all possible questions about the different aspects can be selected for being answered by the subjects. This approach was also used in [7, 8].

Definition 10 (Pers. partial process understandability) *The personal partial process understandability $U_a(p, s, Q_a)$ of aspect a, process p, subject s and questions $Q_a \subseteq Q_{a,max}(p)$ is defined as the fraction of correct answers given by s to the questions Q_a of aspect a about p.*

$$U_a(p, s, Q_a) := \frac{\text{\# correct answers to } Q_a}{|Q_a|} \quad , a \in \{o, c, e, r\} \quad (6)$$

Here again, the different values of personal partial process understandability can be seen as outcomes of a random variable. The expected value of this variable can be estimated according to Definition 11.

Definition 11 (Est. partial process understandability) *The estimated partial process understandability $\widehat{U}_a(p, S, Q_a)$ of aspect a, process p, subjects S and questions Q_a is defined as the average personal partial process understandability of p and Q_a by the subjects of S.*

$$\widehat{U}_a(p, S, Q_a) := \frac{1}{|S|} \sum_{s \in S} U_a(p, s, Q_a) \quad , a \in \{o, c, e, r\} \quad (7)$$

In order to measure the number of actually asked questions Q_a relative to the number of possible questions $Q_{a,max}(p)$ about process p, we define *coverage rate*.

Definition 12 (Coverage rate) *The coverage rate of a set of questions $Q_a \subseteq Q_{a,max}(p)$ about aspect a of process p is defined as*

$$r_a(Q_a, p) := \frac{|Q_a|}{|Q_{a,max}(p)|} \quad , a \in \{o, c, e, r\} \quad . \quad (8)$$

Corollary 3 *The number of different sets of questions $Q_a \subseteq Q_{a,max}(p)$ with $|Q_a| = m$ questions is*

$$\binom{|Q_{a,max}(p)|}{m} \quad . \quad (9)$$

Hypothesis 3 *The different questions of $Q_{a,max}(p)$ are not equally difficult. This has two consequences: (1) For the same coverage rate, one gets different values for estimated partial process understandability depending on the selected questions Q_a. (2) The smaller the coverage rate, the bigger the standard deviation of the different values of estimated partial process understandability for that coverage rate.*

As a consequence, the coverage rate should not be selected too small. Furthermore, the questions for the set Q_a should be chosen randomly in order to minimize the risk of intentionally or unintentionally selecting especially easy or difficult questions when done by a human. The two recommendations shall assure that the estimated partial process understandability does not differ that much from the true value of process understandability.

7 Experimental Evaluation and Results

7.1 Experimental Design

For our experimental evaluation, we used the process depicted in Figure 2. It was presented to the subjects in the same top-to-bottom-style EPC-like notation as in [7, 8].

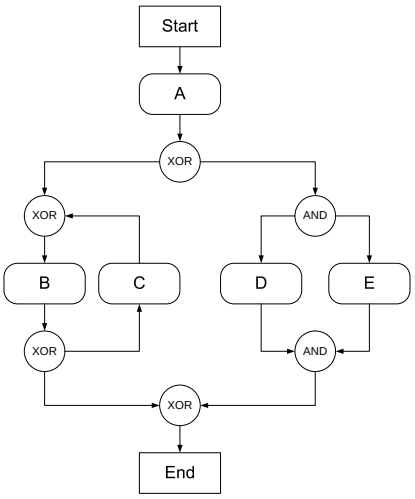


Figure 2: Process used in experiment.

As the process has only five tasks, all $Q_{o,max}(p) = 20$, $Q_{c,max}(p) = Q_{e,max}(p) = 10$ and $Q_{r,max}(p) = 5$ possible questions about the four aspects could have been asked.

We created a questionnaire with two groups (group A: questions about *order* and *repetition*; group B: questions about *concurrency* and *exclusiveness*).

We asked students attending the "Workflow Management" lecture at University Karlsruhe to participate in our experiment. Participation in the experiment was voluntary. Finally, 18 students answered our questionnaire (nine from each group).

7.2 Results

The answers to the questionnaire are given in Table 1 (aspect *order*), Table 2 (aspect *repetition*), Table 3 (aspect *concurrency*) and Table 4 (aspect *exclusiveness*).

The personal process understandability values of the subjects for the four aspects *order* (o), *repetition* (r), *concurrency* (c) and *exclusiveness* (e) are depicted in Figure 3.

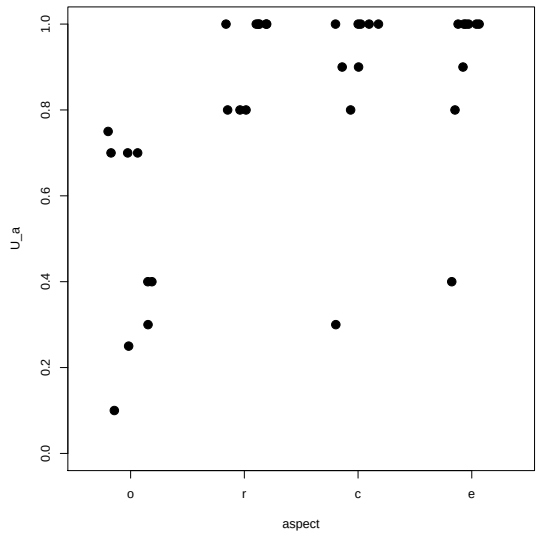


Figure 3: Personal process understandability values for the four aspects.

Regarding Hypothesis 1 In order to test our hypothesis that the personal process understandability values are normally distributed for each aspect, we did a Shapiro-Wilk test [10] for each of the four data sets. For *repetition*, *concurrency* and *exclusiveness*, we had to reject the null-hypothesis that the data is normally distributed ($p \ll 0.05$). Only for *order*, this null-hypothesis could not be rejected on the $\alpha = 0.05$ level.

We can think about the following reasons for not finding a normal distribution for *repetition*, *concurrency* and *exclusiveness*: (1) The process is too "easy". So, most values are near 1.0. As the value range ends there, there cannot exist any bigger values "symmetric" to the values lower than 1.0. (2) The process is too "small". Only five and ten questions were asked respectively. Consequently, personal process understandability values have a "step size" of 0.2 and 0.1 respectively. (3) The number of subjects was too low. We collected only data from nine participants per aspect.

Based on the data about personal process understandability, the estimated process understandability values (together with the standard deviations of the personal process understandability values) were computed (Table 5).

Table 1: Answers given to aspect *order*.

subject	(A,B)	(A,C)	(A,D)	(A,E)	(B,A)	(B,C)	(B,D)	(B,E)	(C,A)	(C,B)	(C,D)	(C,E)	(D,A)	(D,B)	(D,C)	(D,E)	(E,A)	(E,B)	(E,C)	(E,D)	$U_o(p,s)$
solution	$o \sqsubset$	$o \sqsupset$	$o \sqsupset$	$o \sqsupset$	$o \not\sharp$	$o \sqsupset$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \sqsupset$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	
s_1	$o \vee$	$o \vee$	$o \vee$	$o \vee$	$o \not\sharp$	$o \vee$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	0.70
s_3	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \sqsupset$	$o \vee$	$o \vee$	$o \sqsupset$	$o \not\sharp$	$o \vee$	$o \vee$	$o \sqsupset$	$o \vee$	$o \vee$	$o \sqsupset$	$o \vee$	$o \vee$	$o \vee$	$o \sqsupset$	0.10
s_5	$o \vee$	$o \vee$	$o \vee$	$o \vee$	$o \sqsupset$	$o \sqsupset$	$o \sqsupset$	$o \sqsupset$	$o \not\sharp$	$o \sqsupset$	$o \sqsupset$	$o \sqsupset$	$o \not\sharp$	$o \sqsupset$	$o \sqsupset$	$o \not\sharp$	$o \not\sharp$	$o \sqsupset$	$o \sqsupset$	$o \not\sharp$	0.40
s_{11}	$o \vee$	$o \vee$	$o \vee$	$o \vee$	$o \not\sharp$	$o \sqsupset$	$o \not\sharp$	$o \sqsupset$	$o \not\sharp$	$o \not\sharp$	$o \sqsupset$	$o \sqsupset$	$o \not\sharp$	$o \sqsupset$	$o \sqsupset$	$o \not\sharp$	$o \not\sharp$	$o \sqsupset$	$o \sqsupset$	$o \not\sharp$	0.40
s_{35}	$o \vee$	$o \vee$	$o \vee$	$o \vee$	$o \not\sharp$	$o \sqsupset$	$o \not\sharp$	$o \sqsupset$	$o \not\sharp$	$o \not\sharp$	$o \sqsupset$	$o \sqsupset$	$o \not\sharp$	$o \sqsupset$	$o \sqsupset$	$o \not\sharp$	$o \not\sharp$	$o \sqsupset$	$o \sqsupset$	$o \not\sharp$	0.30
s_{51}	$o \vee$	$o \vee$	$o \vee$	$o \vee$	$o \not\sharp$	$o \vee$	$o \not\sharp$	$o \sqsupset$	$o \not\sharp$	$o \not\sharp$	$o \sqsupset$	$o \sqsupset$	$o \not\sharp$	$o \sqsupset$	$o \sqsupset$	$o \not\sharp$	$o \vee$	$o \sqsupset$	$o \sqsupset$	$o \not\sharp$	0.25
s_{53}	$o \vee$	$o \vee$	$o \vee$	$o \vee$	$o \not\sharp$	$o \vee$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \vee$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	0.70
s_{55}	$o \vee$	$o \vee$	$o \vee$	$o \vee$	$o \not\sharp$	$o \sqsupset$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	0.75
s_{57}	$o \vee$	$o \vee$	$o \vee$	$o \vee$	$o \not\sharp$	$o \vee$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	$o \not\sharp$	0.70
correct	0%	0%	0%	0%	100%	44%	56%	44%	89%	11%	44%	44%	89%	44%	44%	89%	78%	44%	44%	89%	

Table 3: Answers given to aspect *concurrency*.

subject	(A,B)	(A,C)	(A,D)	(A,E)	(B,C)	(B,D)	(B,E)	(C,D)	(C,E)	(D,E)	$U_c(p,s)$
solution	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\exists}$	
s_2	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\forall}$	0.9
s_4	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\exists}$	1.0
s_6	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\exists}$	1.0
s_{34}	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\exists}$	1.0
s_{42}	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\forall}$	0.9
s_{50}	$c_{\not\exists}$	$c_{\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\forall}$	$c_{\exists}$	$c_{\exists}$	$c_{\exists}$	$c_{\exists}$	$c_{\not\exists}$	0.3
s_{52}	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\exists}$	-	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\exists}$	0.8
s_{56}	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\exists}$	1.0
s_{60}	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\not\exists}$	$c_{\exists}$	1.0
correct	100%	89%	100%	100%	78%	78%	89%	89%	89%	67%	

Table 4: Answers given to aspect *exclusiveness*.

subject	(A,B)	(A,C)	(A,D)	(A,E)	(B,C)	(B,D)	(B,E)	(C,D)	(C,E)	(D,E)	$U_e(p,s)$
solution	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\exists}$	
s_2	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\exists}$	1.0
s_4	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\exists}$	1.0
s_6	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\exists}$	1.0
s_{34}	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\exists}$	1.0
s_{42}	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\forall}$	0.9
s_{50}	$e_{\forall}$	$e_{\forall}$	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\exists}$	0.8
s_{52}	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\forall}$	$e_{\exists}$	0.4
s_{56}	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\exists}$	1.0
s_{60}	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\not\exists}$	$e_{\exists}$	1.0
correct	78%	78%	89%	89%	89%	100%	100%	100%	89%	89%	

Table 2: Answers given to aspect *repetition*.

subject	A	B	C	D	E	$U_r(p,s)$
solution	$r_{=1}$	r_*	r_*	$r_?$	$r_?$	
s_1	$r_{=1}$	r_*	r_*	$r_?$	$r_?$	1.0
s_3	$r_{=1}$	r_*	r_*	$r_?$	$r_?$	1.0
s_5	$r_{=1}$	r_+	r_*	$r_?$	$r_?$	0.8
s_{11}	$r_{=1}$	r_*	r_*	$r_?$	$r_?$	1.0
s_{35}	$r_{=1}$	r_*	r_*	$r_?$	$r_?$	1.0
s_{51}	$r_{=1}$	r_*	r_*	$r_?$	$r_?$	1.0
s_{53}	$r_{=1}$	r_*	r_*	$r_?$	$r_?$	1.0
s_{55}	$r_{=1}$	$r_?$	r_*	$r_?$	$r_?$	0.8
s_{57}	$r_{=1}$	r_+	r_*	$r_?$	$r_?$	0.8
correct	100%	67%	100%	100%	100%	

Table 5: Estimated process understandability values, standard deviations and 95% confidence intervals for the four aspects.

	order	repetition	concurrency	exclusiveness
$\hat{U}_a(p,S)$	0.478	0.933	0.878	0.900
s.d.	0.240	0.100	0.228	0.200
lower conf. int. bound	0.293	0.866	0.722	0.755
upper conf. int. bound	0.663	0.979	0.989	1.000

We also computed 95% confidence intervals for the expected process understandability values of the four aspects. For *order*, we used the method for estimating confidence intervals for means of normal distributions [9, pp. 446–447]. For the other three aspects, we used the bootstrap approach [1] which does not require normally distributed data. The lower and upper confidence interval bounds are also listed in Table 5.

The estimated process understandability values and the 95% confidence intervals for the four aspects are also depicted graphically in Figure 4.

Regarding Hypothesis 2 For testing our hypothesis that the process understandability values for the four aspects are different, we used Wilcoxon rank-sum tests for independent values (aspects asked for in different experimental groups) [9, pp. 590–597] and Wilcoxon signed-rank tests for paired values (aspects asked for in one single experimental group) [9, pp. 599–603]. Both tests do not require normally distributed data. Only for the combinations *order-repetition*, *order-concurrency* and *order-exclusiveness*, the null-hypothesis (data belongs to same distribution) could be rejected on the $\alpha = 0.05$ level. Here again, a possible reason that the values for *repetition*, *concurrency* and *exclusiveness* are so equal could be that the process is too "small" and "easy" so that no real complicated parts that are differently complicate for the different aspects are included.

Regarding Hypothesis 3 In order to test our hypothesis about partial process understandability, we computed all estimated partial process understandability

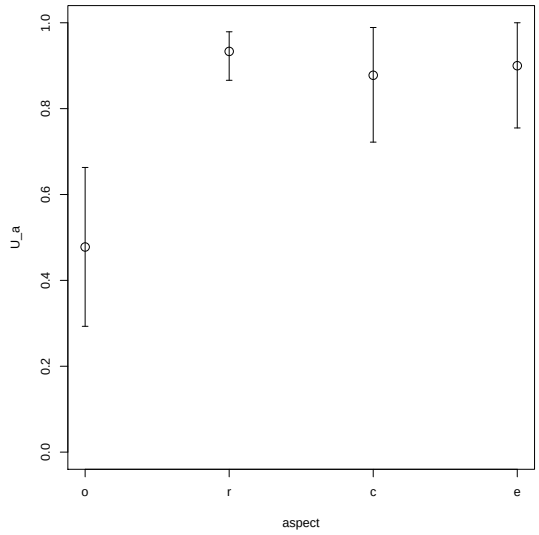


Figure 4: Estimated process understandability values and 95% confidence intervals for the four aspects.

values for the four aspects.

The values depending on the coverage rate are depicted in Figure 5. The dashed horizontal lines are the lower and upper 95% confidence interval bounds for the estimated process understandability values of the four aspects.

In Table 6, the mean estimated partial process understandability, the standard deviation of the estimated partial process understandability values and the rate of values lower and higher than the confidence interval bounds of the four aspects are listed for all different coverage rates.

Table 6 and Figure 5 support our hypothesis—aspect *order* having the strongest effect: For the same coverage rate, many different estimated partial process understandability values exist. The smaller the coverage rate, the higher the standard deviation and the number of values outside the confidence interval.

8 Conclusion and Future Work

In this paper, we gave an overview about the work on measuring process understandability and necessary basics about measurement and prediction systems. We showed points of criticism about the existing measuring approaches—especially the possible lack of reliability and validity of the proposed process understandability metrics. We gave concrete and detailed definitions for measuring process understandability and formalized our points of criticism as hypotheses that we subsequently examined in an experiment.

The experiment supports our hypotheses that different aspects of process understandability can be differently complicated and that asking only a little part of the set of possible questions can cause values for process understandability differing very much from the real value.

Consequently, all different aspects of process understandability have to be measured to get an overall value. The coverage rate of asked questions must not be too small. The questions should be selected randomly to minimize the risk of choosing especially easy or difficult questions.

For future work, we suggest additional experiments with larger processes and more subjects to check our results and to test whether the effect of different values for the different aspects of process understandability becomes even more obvious. Furthermore, it should be examined whether there are other aspects of understandability not identified so far. The selection of proper coverage rates minimizing the measuring effort *and* the differences from the real process understandability value should also be investigated.

Acknowledgment

The authors want to thank Agnes Koschmider for the possibility to conduct the experiment in the "Workflow Management" lecture as well as the participating students.

References

[1] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

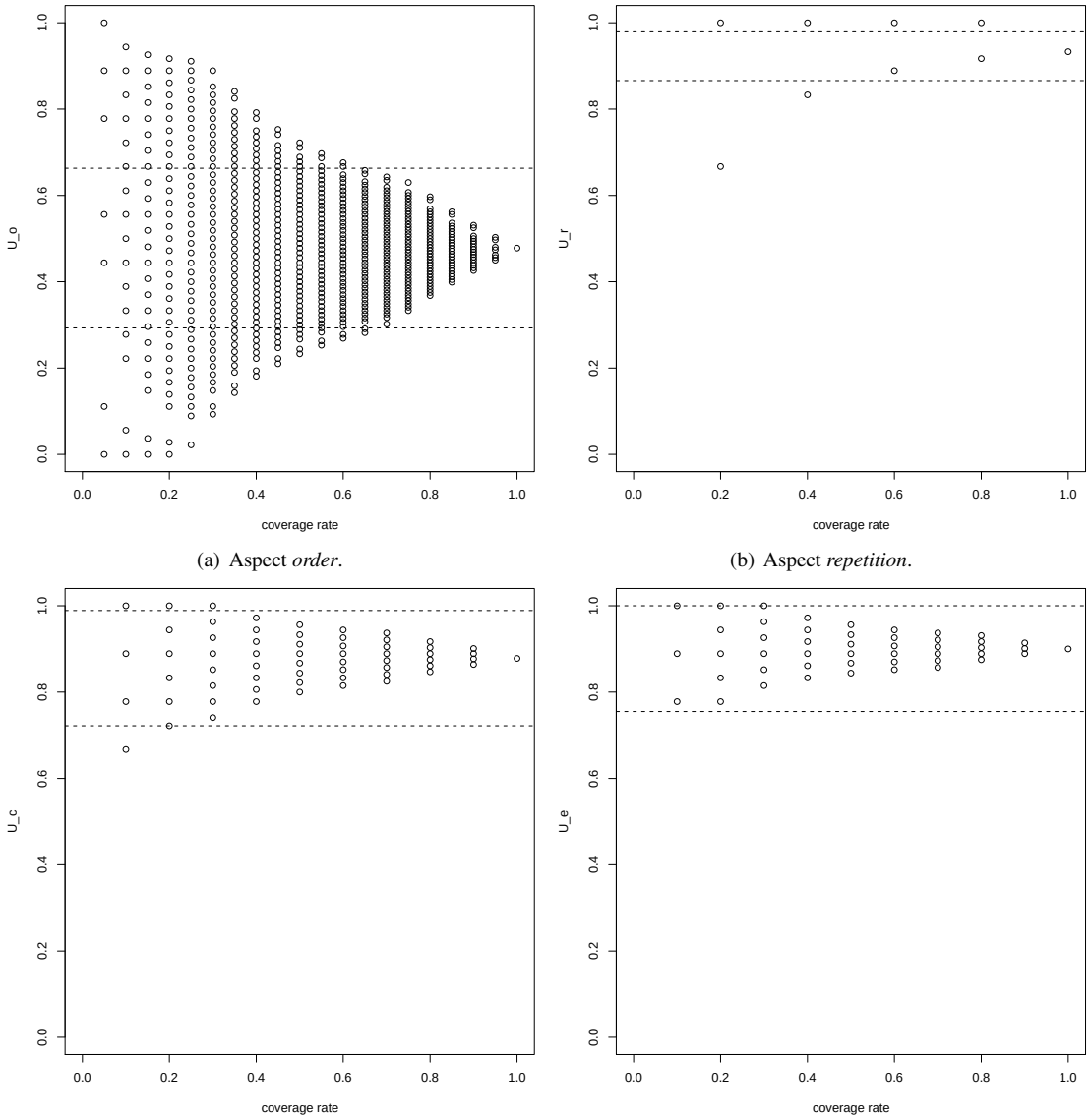
[2] Norman E. Fenton and Shari Lawrence Pfleeger. *Software Metrics: A Rigorous and Practical Approach*. International Thomson Computer Press, 2nd edition, 1996.

[3] Andrew Gemino and Yair Wand. Evaluating modeling techniques based on models of learning. *Communications of the ACM*, 46(10):79–84, 2003.

[4] Stephen H. Kan. *Metrics and Models in Software Quality Engineering*. Addison-Wesley, 2nd edition, 2002.

[5] Richard E. Mayer. Models for understanding. *Review of Educational Research*, 59(1):43–64, 1989.

[6] Joachim Melcher and Detlef Seese. Process measurement: Insights from software measurement on measuring process complexity, quality and performance. Research report, Universität Karlsruhe (TH), Institut AIFB, 2008. `http://digbib.ubka.uni-karlsruhe.de/volltexte/1000009225`.



(a) Aspect *order*.

(b) Aspect *repetition*.

(c) Aspect *concurrency*.

(d) Aspect *exclusiveness*.

Figure 5: Estimated partial process understandability values of the four aspects depending on coverage rate.

[7] Jan Mendling, Hajo A. Reijers, and Jorge Cardoso. What makes process models understandable? In Gustavo Alonso, Peter Dadam, and Michael Rosemann, editors, *Business Process Management: Proceedings of the 5th International Conference BPM 2007*, volume 4714 of *LNCS*, pages 48–63, 2007.

[8] Jan Mendling and Mark Strembeck. Influence factors of understanding business process models. In Witold Abramowicz and Dieter Fensel, editors, *Business Information Systems: Proceedings of the 11th International Conference BIS 2008*, volume 7 of *Lecture Notes in Business Information Processing*, pages 142–153, 2008.

[9] Michael J. Panik. *Advanced Statistics from an Elementary Point of View*. Elsevier Academic Press, 2005.

[10] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4):591–611, 1965.

[11] Irene Vanderfeesten, Hajo A. Reijers, Jan Mendling, Wil M. P. van der Aalst, and Jorge Cardoso. On a quest for good process models: The cross-connectivity metric. In Zohra Bellahsène and Michel Léonard, editors, *Advanced Information Systems Engineering: Proceedings of the 20th International Conference CAiSE 2008*, volume 5074 of *LNCS*, pages 480–494, 2008.

Table 6: Data about estimated partial process understandability values of the four aspects.

(a) Aspect *order*.

# questions	cov. rate	mean	s. d.	rate lower	rate higher
1	0.05	0.478	0.333	25.0%	30.0%
2	0.10	0.478	0.224	28.4%	32.1%
3	0.15	0.478	0.177	9.3%	14.6%
4	0.20	0.478	0.149	10.6%	14.4%
5	0.25	0.478	0.129	10.5%	7.3%
6	0.30	0.478	0.114	3.7%	6.5%
7	0.35	0.478	0.101	3.7%	3.1%
8	0.40	0.478	0.091	2.8%	2.5%
9	0.45	0.478	0.082	0.8%	1.0%
10	0.50	0.478	0.074	0.7%	0.7%
11	0.55	0.478	0.067	0.3%	0.1%
12	0.60	0.478	0.061	0.0%	0.0%
13	0.65	0.478	0.055	0.0%	0.0%
14	0.70	0.478	0.049	0.0%	0.0%
15	0.75	0.478	0.043	0.0%	0.0%
16	0.80	0.478	0.037	0.0%	0.0%
17	0.85	0.478	0.031	0.0%	0.0%
18	0.90	0.478	0.025	0.0%	0.0%
19	0.95	0.478	0.018	0.0%	0.0%
20	1.00	0.478	–	0.0%	0.0%

(b) Aspect *repetition*.

# questions	cov. rate	mean	s. d.	rate lower	rate higher
1	0.2	0.933	0.149	20.0%	80.0%
2	0.4	0.933	0.086	40.0%	60.0%
3	0.6	0.933	0.057	0.0%	40.8%
4	0.8	0.933	0.037	0.0%	20.0%
5	1.0	0.933	–	0.0%	0.0%

(c) Aspect *concurrency*.

# questions	cov. rate	mean	s. d.	rate lower	rate higher
1	0.1	0.878	0.110	10.0%	30.0%
2	0.2	0.878	0.071	0.0%	6.7%
3	0.3	0.878	0.054	0.0%	0.8%
4	0.4	0.878	0.043	0.0%	0.0%
5	0.5	0.878	0.035	0.0%	0.0%
6	0.6	0.878	0.029	0.0%	0.0%
7	0.7	0.878	0.023	0.0%	0.0%
8	0.8	0.878	0.018	0.0%	0.0%
9	0.9	0.878	0.012	0.0%	0.0%
10	1.0	0.878	–	0.0%	0.0%

(d) Aspect *exclusiveness*.

# questions	cov. rate	mean	s. d.	rate lower	rate higher
1	0.1	0.900	0.082	0.0%	0.0%
2	0.2	0.900	0.052	0.0%	0.0%
3	0.3	0.900	0.040	0.0%	0.8%
4	0.4	0.900	0.032	0.0%	0.0%
5	0.5	0.900	0.026	0.0%	0.0%
6	0.6	0.900	0.021	0.0%	0.0%
7	0.7	0.900	0.017	0.0%	0.0%
8	0.8	0.900	0.013	0.0%	0.0%
9	0.9	0.900	0.009	0.0%	0.0%
10	1.0	0.900	–	0.0%	0.0%