

On Measuring the Understandability of Process Models (Experimental Results)

Joachim Melcher¹, Jan Mendling², Hajo A. Reijers³ and Detlef Seese¹

¹ Institut AIFB, Universität Karlsruhe (TH)
76128 Karlsruhe, Germany
{melcher|seese}@aifb.uni-karlsruhe.de

² Institut für Wirtschaftsinformatik, Humboldt-Universität zu Berlin
10099 Berlin, Germany
jan.mendling@wiwi.hu-berlin.de

³ School of Industrial Engineering, Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
h.a.reijers@tue.nl

July 10, 2009

Abstract

Much efforts are aimed at unveiling the factors that influence a person's comprehension of a business process model. While various potential factors have been proposed and studied in an experimental setting, little attention is being paid to the measuring of the structural process understanding itself.

In a previous paper, Melcher and Seese identified possible problems with metric reliability and validity for these few existing measurement approaches. In this previous paper, they defined new metrics for structural process understandability inspired by existing work. Conducting a small experiment, they could support some hypotheses about their behavior.

In this paper, we present the results of a larger experiment involving 178 students from three universities. We could reconfirm the findings of the first experiment and got some encouraging additional findings supporting our hypotheses: Different aspects of struc-

tural process understandability can be complicated in varying degrees for a process model and asking only some few questions about a process model can cause values for structural process understandability differing very much from the real value.

This paper provides various recommendations to properly measure structural process understandability which should be considered in future work.

1 Introduction

Process models are often used as a communication vehicle among people, for example to clarify how a new information system is expected to support operations or to explain a new employee's duties in a particular field. These examples motivate the importance of process models being clearly understandable. It has been suggested that various factors play a role in the sense-making of a process model, like a reader's expertise [13], the used modeling notation [15], the visual layout of the model [11, 14] and the structural attributes of the process model [8, 17].

Motivated by the research in software metrics, several papers proposing numerous process metrics for measuring internal (often structural) process model attributes have been proposed (see, e.g., [5] for an overview). Integrated in *validated* prediction systems, these process metrics can be used to predict the values of external process model attributes (e.g., error-



This work is licensed under the Creative Commons Attribution-NonCommercial-No Derivative Works 3.0 Germany License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/3.0/de/>.

proneness, time, costs) even before a process model has been implemented or for external attributes that are measurable only in a very laborious manner.

One very important external attribute is structural process understandability by involved humans (e.g., process designers, process analysts, process implementers or people executing a process). Understandability influences other quality aspects of process models 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” [9, p. 48].

While certain insights can be derived from relevant theories, for example [3], it is clear that empirical research is required to substantiate and adapt these for the process modeling domain. A common set-up for that kind of research, and one that has been applied in earlier work [9, 10], is to proceed as follows. One selects an independent variable, for example, a process reader’s familiarity with a notation, and aims to study the relation between its variation and that of a particular measure for the reader’s structural understanding of a model, the dependent variable. Typically, respondents are asked to study one or more process models and then provide answers to questions related to these. The answers can then be used to quantify their structural comprehension of the process models in question.

For examining structural process understandability and validating appropriate prediction systems, we first have to quantify structural process understandability. So, we have to find a proper structural process understandability metric fulfilling the reliability and validity requirements for metrics.

In [6, 7], Melcher and Seese show problems with reliability and validity of metrics for structural process understandability proposed so far. They, give concrete and detailed definitions for measuring structural process understandability exceeding those in existing publications. Conducting a rather small experiment, they show some effects of using these metrics.

In this paper, we present the results of a much larger second experiment using a more realistic process model and involving 178 students from three universities.

The remainder of this paper is organized as follows: In Section 2, we present related work on measuring structural process understandability. Important basics about measurement and prediction systems are shown in Section 3. Our definitions for measuring structural process understandability and hypotheses about some effects of measurement are shown in Section 4. In Section 5, an experiment for examining these hypotheses is presented. The paper gives a conclusion and presents possible future work (Section 6).

2 Related Work

In [9], Mendling *et al.* searched for possible relations between personal and process specific (structural) properties and structural process understandability.

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

As operationalization of structural 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 model. The sum of correct answers (at most nine) gives the SCORE value.

In addition to the goals of [9], Mendling and Strembeck examined also the influence of content related factors on structural process understandability in [10].

For that purpose, they designed an online questionnaire that was answered by 42 students and practitioners. Six process models 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 process models were depicted in the same notation as in [9]. For each process model, six yes/no questions about process model 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 process model (at most 36) and served as an operationalization of structural understandability to a person.

In [6, 7], Melcher and Seese give an overview about existing publications trying to measure structural process understandability. They introduce basics of measurement and prediction systems (see Section 3). Based on these theoretic concepts, they show a lack of reliability and validity of the proposed metrics.

Inspired by existing work, they define new metrics for structural process understandability and formulate hypotheses about effects of measuring structural process understandability. Conducting a rather small experiment, they got some encouraging findings supporting these hypotheses.

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 [5], Melcher and Seese show that these theoretical basics are also essential for process measurement—even so they 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 [5], Melcher and Seese show how the idea of prediction systems can be transferred to process measurement (see Figure 1):

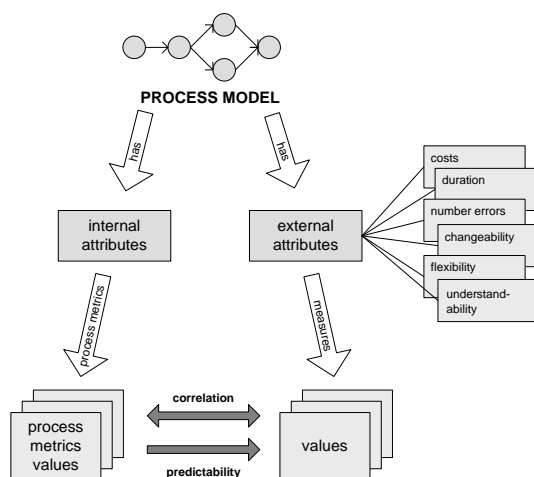


Figure 1: Prediction systems adapted to process measurement.

A process model has *internal* and *external* attributes. Internal attributes are those that can be measured purely in terms of the process model 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 model relates to its environment [2, p. 74]. Examples are costs, time, number of errors and—especially important for this paper—(structural) 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 Measuring Structural Process Understandability

The process understandability¹ metrics, other definitions and hypotheses presented in this section (except from those in Subsection 4.3) are the same as in [6, 7].

4.1 Aspects of Process Understandability

It is important to cover the different aspects of structural process understandability in order to meet the con-

¹For the sake of simplicity, we use the term “process understandability” instead of “*structural* process understandability” in the rest of this paper.

tent validity requirement for metrics. Our definitions formalize the aspects *concurrency*, *exclusiveness*, *order* and *repetition* which are identified by Mendling *et al.* in [9, p. 52]. While these notions are thought to cover a broad array of understandability aspects, we do not deny the possible existence of other aspects.

Our definitions are based on the concept of an activity period.

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.*

Using this concept, we can define relations for the four aspects of process understandability we mentioned.

Definition 4 (Concurrency) *For the questions about task concurrency, the relations $c_{\neq}, c_{\exists}, c_{\forall} \subseteq T \times T$ with the following meanings are used.*

$(t_1, t_2) \in c_{\neq} \Leftrightarrow$ *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 5 (Exclusiveness) *For the questions about task exclusiveness, the relations $e_{\neq}, e_{\exists}, e_{\forall} \subseteq T \times T$ with the following meanings are used.*

$(t_1, t_2) \in e_{\neq} \Leftrightarrow$ *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 6 (Order) *For the questions about task order, the relations $o_{\neq}, o_{\exists}, o_{\forall} \subseteq T \times T$ with the following meanings are used.*

$(t_1, t_2) \in o_{\neq} \Leftrightarrow$ *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 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.*

We constructed these definitions 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 $c_{\neq}, c_{\exists}, c_{\forall}$ and $e_{\neq}, e_{\exists}, e_{\forall}$ 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 model: 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 $a \in \{c, e, o, r\}$ about a process model p with n tasks is*

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

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

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

As one can see, the maximum number of questions for *concurrency*, *exclusiveness* and *order* 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 model 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 .

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

Hypothesis 1 *The personal process understandability metric values $U_a(p, s_i)$ of a process model 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 as follows.

Definition 9 (Estimated process understandability)

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

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

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 $\hat{U}_a(p, S)$ of a process model p .*

Consequently, it is important to measure at least all of these aspects to achieve an insight into one’s “overall understandability” of a model.

4.2 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 [9, 10].

Definition 10 (Pers. partial process understandability)

The personal partial process understandability $U_a(p, s, Q_a)$ of aspect a , process model 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 .

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

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 $\hat{U}_a(p, S, Q_a)$ of aspect a , process model 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 .

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

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 model 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 model p is defined as*

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

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

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

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. The two recommendations shall assure that the estimated partial process understandability does not differ that much from the true value of process understandability.

4.3 Process Understandability Using Virtual Subjects

As the number of possible questions soon becomes so high for larger process models (see Corollary 2), not all of them can be asked to one single subject. Besides using partial process understandability (Subsection 4.2), there is a second approach—virtual subjects—which is based on the following hypothesis.

Hypothesis 4 *Randomly dividing a set of questions answered by a group of subjects into two subsets of approximately same size results in a strong correlation between the rates of correct answers given by the same subject to the questions of the two subsets.*

Roughly speaking that means that a subject with good results for one subset of questions will also be good for the second subset. This is used by us in inverse direction in order to “construct” a new virtual subject’s answers out of the answers given by several real subjects:

The set of all possible questions of one aspect is divided into different subsets which are each answered by different groups of subjects. Afterwards, in each group the subjects are ordered by their personal partial process understandability values. Now, new virtual subjects are “created” by combining the answers of one subject from each group. For this step, the best subjects from each group are combined to the best new virtual subject, the second best subjects to the second best new virtual subject—and so on.

Using the answers of these so “constructed” virtual subjects, (virtual) personal process understandability and (virtual) estimated process understandability can be computed as defined in Definitions 8 and 9 respectively.

5 Experimental Evaluation and Results

5.1 Experimental Design

For our experimental evaluation, we used the process model depicted in Figure 2. We used the same top-to-bottom-style layout as in [6, 7, 9, 10].

As the process model has 12 tasks, the number of possible questions about the four aspects are $|Q_{c,max}(p)| = |Q_{e,max}(p)| = 66$, $|Q_{o,max}(p)| = 132$ and $|Q_{r,max}(p)| = 12$.

As the number of possible questions per aspect is too high (except for aspect *repetition*), we could not asked them all to one subject. Instead, we divided the questions for *concurrency*, *exclusiveness* and *order* into different subsets. So, we created a questionnaire with nine groups (groups 1 to 4: questions about *order* [o1–o4];

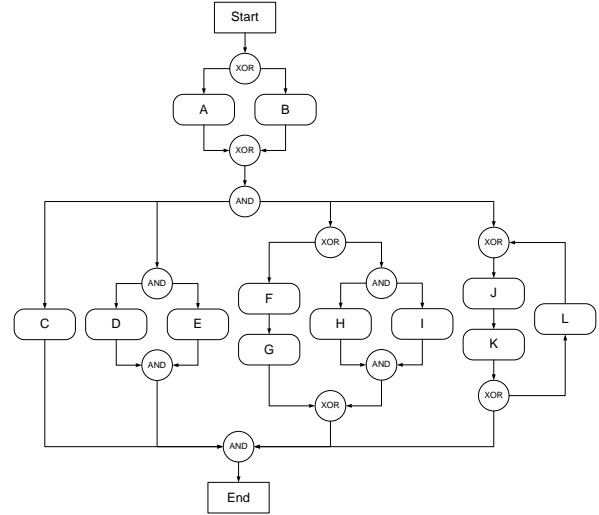


Figure 2: Process model used in experiment.

groups 5 to 8: questions about *concurrency* [c5–c8] and *exclusiveness* [e5–e8]; group 9: questions about *repetition* [r9])—resulting in 13 data sets. In each group, we asked 33 questions (group 9 was filled by 21 “dummy questions”). The detailed assignment of the questions to the groups is shown in Tables 1–9.

We asked students attending courses on workflow management at Humboldt-Universität zu Berlin, Eindhoven University of Technology and Universität Karlsruhe (TH) to participate in the experiment. Participation was voluntary. Students from Berlin and Eindhoven got a bonus for their final exam—students from Karlsruhe could use it as training for their exam. Altogether, 178 students completed the questionnaire. The participants were randomly assigned to one group of the questionnaire.

5.2 Results

The answers to the questionnaire are given in Tables 1–9.

5.2.1 Results Concerning Single Questions

In this paragraph, we look at the rate of correctly answered questions about the four aspects. These values are given in Table 10 and as histograms in Figure 3.

As one can see, aspect *order* has a quite different behavior compared to the other three aspects. While those have narrow peaks near the rate of 1.0, the values for aspect *order* are more spread over the whole interval with the peak near 0.6. So, most questions of aspect *order* seem to be more difficult to be answered by the subjects than those of the other aspects.

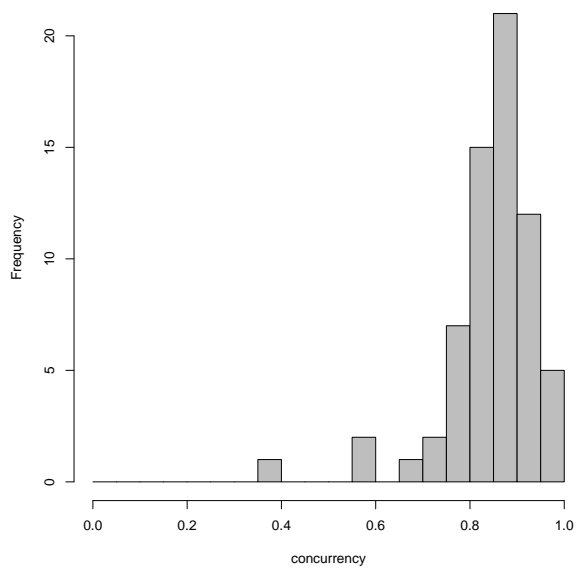
Next, we analyzed whether a connection between the rates of correct answers to questions about *concurrency*

Table 1: Answers given in group 1.

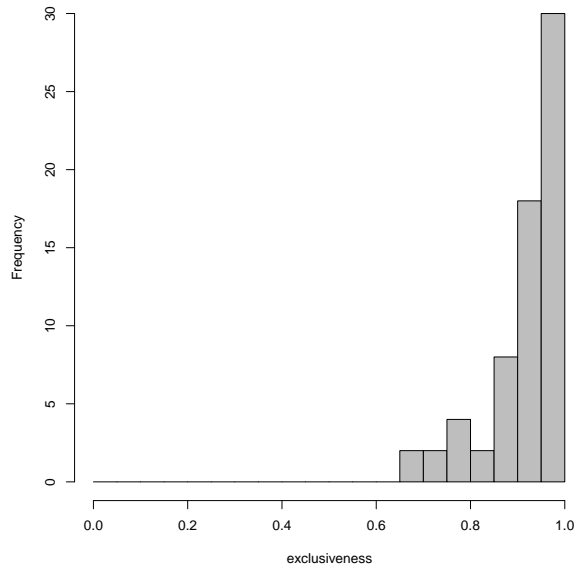
subject solution	qo (A.B)	qo (A.F)	qo (A.J)	qo (B.C)	qo (B.G)	qo (B.K)	qo (C.D)	qo (C.H)	qo (C.L)	qo (D.E)	qo (D.I)	qo (E.A)	qo (E.F)	qo (E.J)	qo (F.B)	qo (F.G)	qo (F.K)	qo (G.C)	qo (G.H)	qo (G.L)	qo (H.D)	qo (H.I)	qo (L.A)	qo (L.B)	qo (L.J)	qo (L.F)	qo (L.K)	qo (K.C)	qo (K.G)	qo (K.L)	qo (L.D)	qo (L.H)	$U_o(p, s, Q_o)$	
	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO		EO
8001	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.79	
8002	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.48	
8003	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.67	
8004	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.61	
8005	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.64	
8006	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.58	
8007	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.27	
8008	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.55	
8009	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.79	
8010	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.42	
8011	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.55	
8012	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.52	
8013	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.58	
8014	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.52	
8015	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.67	
8016	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.70	
8017	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.24	
8018	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.85
correct	44%	61%	50%	67%	61%	61%	72%	83%	83%	28%	50%	50%	33%	56%	50%	6%	89%	78%	50%	100%	44%	83%	67%	67%	33%	56%	33%	94%	72%	89%	33%	61%	33%	

Table 2: Answers given in group 2.

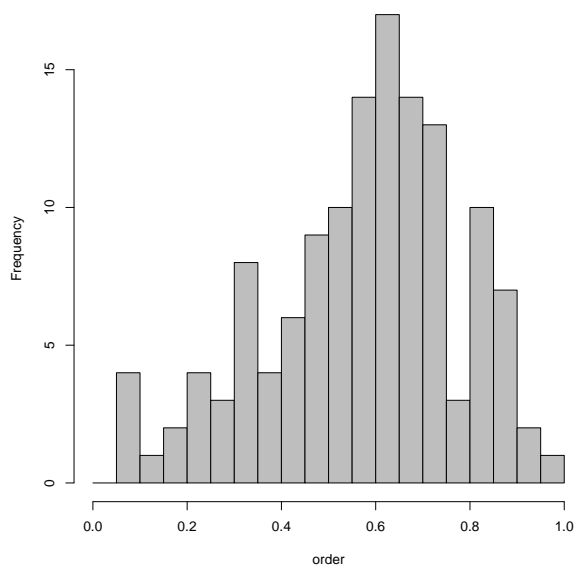
subject solution	qo (A.C)	qo (A.G)	qo (A.K)	qo (B.D)	qo (B.H)	qo (B.L)	qo (C.F)	qo (C.I)	qo (D.A)	qo (D.P)	qo (D.J)	qo (E.B)	qo (E.G)	qo (E.K)	qo (F.C)	qo (F.H)	qo (F.L)	qo (G.D)	qo (G.I)	qo (H.A)	qo (H.E)	qo (H.J)	qo (I.B)	qo (I.F)	qo (I.K)	qo (J.C)	qo (J.G)	qo (J.L)	qo (K.D)	qo (K.H)	qo (L.A)	qo (L.E)	qo (L.I)	$U_o(p, s, Q_o)$			
	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO		EO	EO	EO
8019	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.58	
8020	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.45	
8021	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.76	
8022	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.94	
8023	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.36	
8024	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.64	
8025	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.42	
8026	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.91	
8027	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.52	
8028	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.58	
8029	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.24	
8030	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.58	
8031	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.38	
8032	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.52	
8033	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.70	
8034	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.55	
8035	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.88	
8036	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.45	
8037	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.76	
8038	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	EO	0.76
correct	55%	60%	60%	60%	60%	60%	80%	40%	50%	35%	70%	45%	85%	85%	65%	60%	90%	70%	55%	65%	70%	50%	50%	60%	85%	70%	85%	15%	70%	40%	55%	65%	35%				



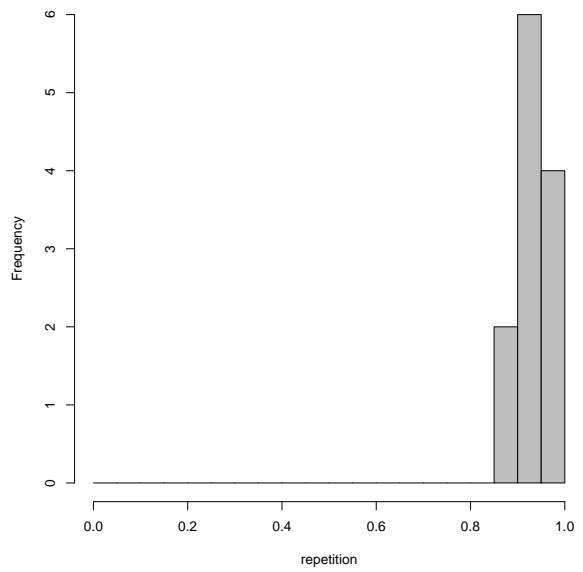
(a) Aspect *concurrency*.



(b) Aspect *exclusiveness*.



(c) Aspect *order*.



(d) Aspect *repetition*.

Figure 3: Histograms for rates of correct answers for the four aspects.

Table 9: Answers given in group 9.

subject	$q_r(A)$	$q_r(B)$	$q_r(C)$	$q_r(D)$	$q_r(E)$	$q_r(F)$	$q_r(G)$	$q_r(H)$	$q_r(I)$	$q_r(J)$	$q_r(K)$	$q_r(L)$	$\bar{U}_r(p, s)$
solution	$r_?$	$r_?$	$r_{=1}$	$r_{=1}$	$r_{=1}$	$r_?$	$r_?$	$r_?$	$r_?$	r_+	r_+	r_*	
s_{164}	$r_?$	$r_?$	$r_{=1}$	$r_{=1}$	$r_{=1}$	$r_?$	$r_?$	$r_?$	$r_?$	r_+	r_+	r_*	1.00
s_{165}	$r_?$	$r_?$	$r_{=1}$	$r_{=1}$	$r_{=1}$	$r_?$	$r_?$	$r_?$	$r_?$	r_+	r_+	r_*	1.00
s_{166}	$r_?$	$r_?$	$r_{=1}$	$r_{=1}$	$r_{=1}$	$r_?$	$r_?$	$r_?$	$r_?$	r_+	r_+	r_*	1.00
s_{167}	$r_?$	$r_?$	$r_{=1}$	$r_{=1}$	$r_{=1}$	$r_?$	$r_?$	$r_?$	$r_?$	r_+	r_+	r_*	1.00
s_{168}	r_*	$r_?$	$r_{=1}$	$r_{=1}$	$r_{=1}$	$r_?$	r_*	$r_?$	$r_?$	r_+	r_+	r_*	0.83
s_{169}	$r_?$	$r_?$	$r_{=1}$	$r_{=1}$	$r_{=1}$	$r_?$	$r_?$	$r_?$	$r_?$	r_+	r_+	r_*	1.00
s_{170}	$r_?$	$r_?$	$r_{=1}$	$r_{=1}$	$r_{=1}$	$r_?$	$r_?$	$r_?$	$r_?$	r_+	r_*	r_*	0.92
s_{171}	$r_?$	$r_?$	$r_{=1}$	$r_{=1}$	$r_{=1}$	$r_?$	$r_?$	$r_?$	$r_?$	r_+	r_+	r_*	1.00
s_{172}	$r_?$	$r_?$	$r_{=1}$	$r_{=1}$	$r_{=1}$	$r_?$	$r_?$	$r_?$	$r_?$	r_+	r_+	r_*	1.00
s_{173}	$r_?$	$r_?$	$r_{=1}$	$r_{=1}$	$r_{=1}$	$r_?$	$r_?$	$r_?$	$r_?$	r_+	r_+	r_*	1.00
s_{174}	$r_?$	$r_?$	r_+	r_*	r_*	$r_{=1}$	$r_{=1}$	$r_?$	$r_{=1}$	r_+	$r_?$	r_*	0.42
s_{175}	$r_?$	$r_?$	$r_{=1}$	$r_{=1}$	$r_{=1}$	$r_?$	$r_?$	$r_?$	$r_?$	r_+	r_+	r_*	1.00
s_{176}	$r_?$	$r_?$	$r_{=1}$	$r_{=1}$	$r_{=1}$	$r_?$	$r_?$	$r_?$	$r_?$	r_+	r_+	r_*	1.00
s_{177}	$r_?$	$r_?$	$r_{=1}$	$r_{=1}$	$r_{=1}$	$r_?$	$r_?$	$r_?$	$r_?$	r_+	r_+	r_*	1.00
s_{178}	$r_?$	$r_?$	$r_{=1}$	$r_{=1}$	$r_{=1}$	$r_?$	$r_?$	$r_?$	$r_?$	r_+	r_+	r_*	1.00
correct	93%	100%	93%	93%	93%	93%	87%	100%	93%	100%	87%	100%	

and *exclusiveness* for the same pair of tasks exist. As both aspects deal with the execution of task pairs during a process instance execution, such a connection is imaginable. The single value pairs are depicted in Figure 4(a). As the Spearman correlation [12, pp. 42–45] is only 0.465, there is no strong connection.

We did the same analysis for aspect *order* and *order* in revers ordering. The single value pairs are depicted in Figure 4(b). Here, the Spearman correlation is -0.209. So, knowing the rate of correct answers to question $q_o(t_1, t_2)$, no prediction for $q_o(t_2, t_1)$ can be given.

5.2.2 Results Concerning (Partial) Process Understandability

The personal (partial) understandability values of the subjects of the nine groups are depicted in Figures 5 and 6.

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 [16] for each of the 13 data sets. Only for $o1$, $o2$ and $o4$, the null-hypothesis that the data is normally distributed could not be rejected on the $\alpha = 0.05$ level. For the rest, we had to reject the null-hypothesis ($o3$: $p = 0.037$; $c5$: $p = 0.035$; all others: $p \ll 0.05$).

Based on the data about personal (partial) process understandability, the estimated (partial) process understandability values (together with the standard deviations of the personal (partial) process understandability values) were computed (Table 11).

We also computed 95% confidence intervals for the estimated (partial) process understandability values of the 13 data sets. For $o1$, $o2$ and $o4$, we used the method for estimating confidence intervals for means of normal

distributions [12, pp. 446–447]. For the other ten data sets, 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 11.

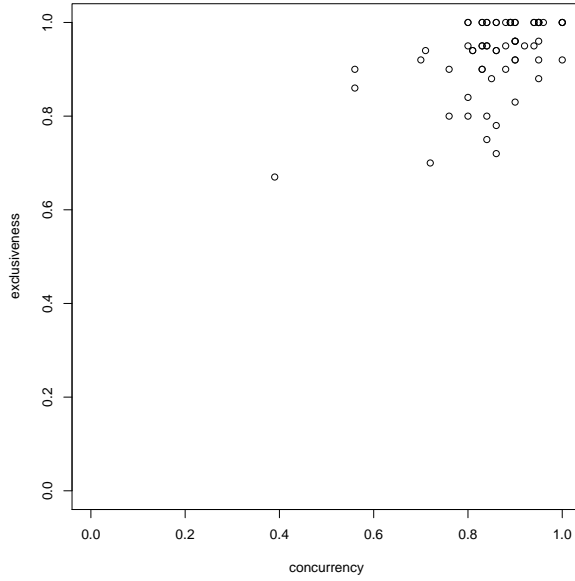
The estimated (partial) process understandability values and the 95% confidence intervals for the 13 data sets are also depicted graphically in Figure 7.

Finally, we analyzed whether the distributions of the personal partial process understandability values of the four data sets of each of the aspects *concurrency*, *exclusiveness* and *order* are the same. For that purpose, we did a Kruskal-Wallis rank sum test for each of these three aspects. The null-hypothesis (same distribution) could not be rejected on the $\alpha = 0.05$ level for all of them. So, the difficulty of the four subsets of questions for each of these aspects seems to be quite equivalent.

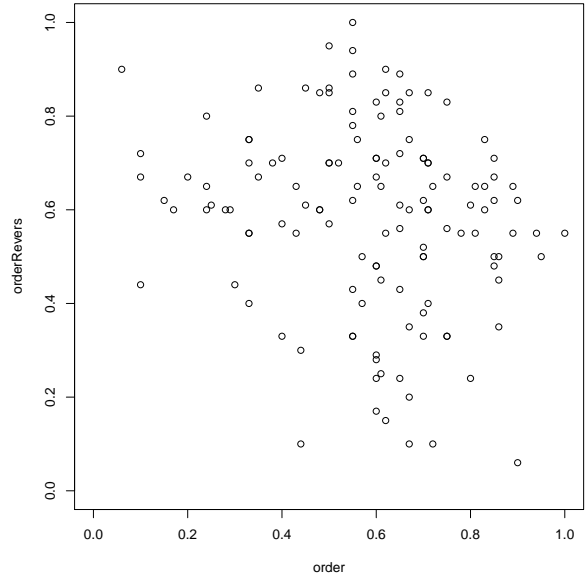
5.2.3 Results Concerning Virtual Subjects Approach

As the process model used in the experiment is so large that the high number of questions could not be asked to single subjects, the questions were divided into different subsets (see Subsection 5.1) for later use of the virtual subjects approach (see Subsection 4.3).

Regarding Hypothesis 4 In order to show that this approach is legitimate, the following test was conducted: Using the data about the aspects *concurrency*, *exclusiveness* and *order* from the experiment in [6, 7], the questions for each aspect were randomly divided into two halves of the same size simulating two groups of questions which could be answered by two different groups of subjects. In the next step, the Spearman rank correlation coefficient between the personal partial process model understandability values from the two halves was computed. This was repeated 5,000 times for each aspect.



(a) Rates of correct answers *concurrency/exclusiveness*.



(b) Rates of correct answers *order/order* in revers order.

Figure 4: Scatter plots with rates of correct answers.

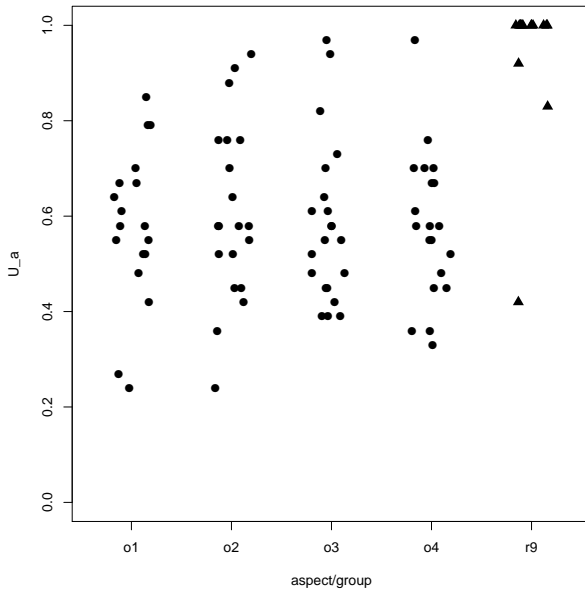


Figure 5: Personal (partial) process understandability values for aspects *order* and *repetition*.

The corresponding empirical cumulative distribution functions are depicted in Figure 8. The medians were 0.714 (*concurrency*), 0.818 (*exclusiveness*) and 0.933 (*order*). So, the approach seems to be legitimate.

The resulting virtual subjects are listed in Table 12.

In the remainder of this chapter, the resulting virtual personal process model understandability values are denoted as $U_a^*(p, s)$, the virtual estimated process model understandability values as $\hat{U}_a^*(p, S)$ and the vir-

tual estimated partial process model understandability as $\hat{U}_a^*(p, S, Q_a)$ ($a \in \{c, e, o, r\}$).

5.2.4 Results Concerning (Virtual) Process Understandability

The (virtual) personal process understandability values of the (virtual) subjects for the four aspects *concurrency*, *exclusiveness*, *order* and *repetition*, are depicted in Figure 9(a).

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 for each of the four data sets. For *concurrency*, *exclusiveness* and *repetition*, we had to reject the null-hypothesis that the data is normally distributed (*concurrency*: $p = 0.023$; all others: $p \ll 0.05$). Only for *order*, this null-hypothesis could not be rejected on the $\alpha = 0.05$ level.

Based on the four data sets, the (virtual) estimated process understandability values (together with the standard deviations of the (virtual) personal process understandability values) were computed (Table 13).

We also computed 95% confidence intervals for the (virtual) estimated process understandability values of the four aspects. For *order*, we used the method for estimating confidence intervals for means of normal distributions [12, 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 13.

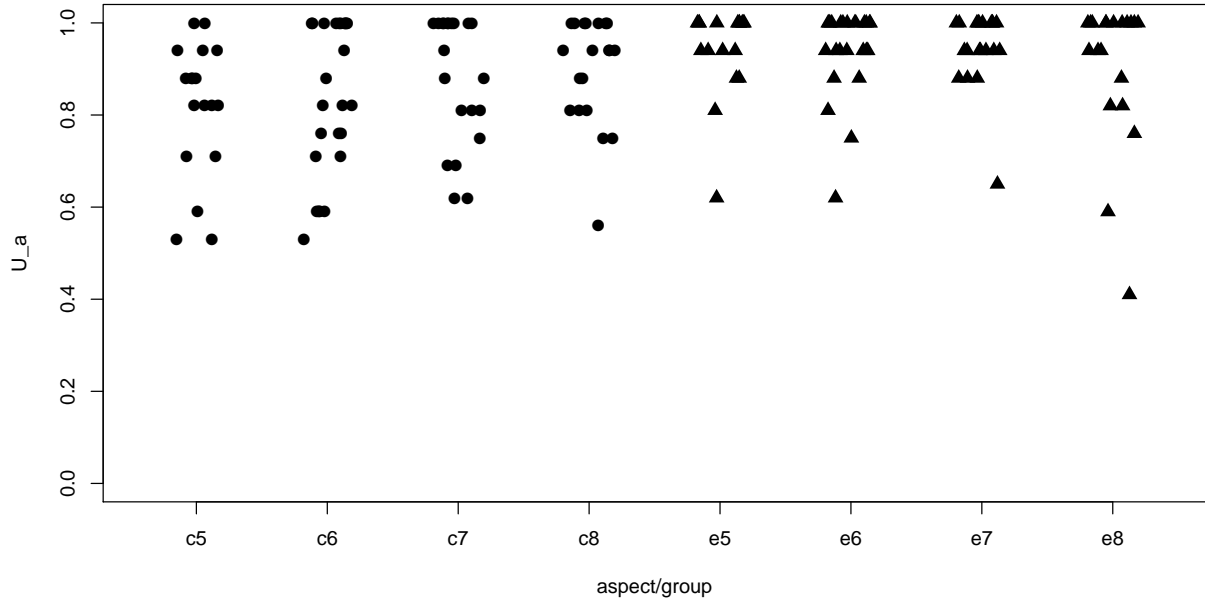


Figure 6: Personal partial process understandability values for aspects *concurrency* and *exclusiveness*.

The (virtual) estimated process understandability values and the 95% confidence intervals for the four aspects are also depicted graphically in Figure 9(b).

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 [12, pp. 590–597]. This test does not require normally distributed data. Only for the combination *exclusiveness-repetition*, the null-hypothesis (data belongs to same distribution) could not be rejected on the $\alpha = 0.05$ level ($p = 0.110$).

5.2.5 Results Concerning (Virtual) Partial Process Understandability

Regarding Hypothesis 3 In order to test our hypothesis about partial process understandability, we computed all (virtual) estimated partial process understandability values for the four aspects. For *concurrency*, *exclusiveness* and *order*, we used the data of the virtual subjects constructed in Subsubsection 5.2.3.

The values depending on the coverage rate are depicted in Figure 11. The dashed horizontal lines are the lower and upper 95% confidence interval bounds for the (virtual) estimated process understandability values of the four aspects.

Because of the “combinatoric explosion” (cf. Corollary 2)², we had to use a probabilistic algorithm for

²The highest number of possible subsets exists for aspect *order* and coverage rate 0.5. Here, $\binom{132}{66} \approx 3.8 \times 10^{38}$ different subsets exist.

these plots: For each analyzed coverage rate, we randomly selected 1,000,000 (for aspects *concurrency* and *exclusiveness*) and 5,000,000 (for aspect *order*) subsets of questions, respectively. Exact values could only be computed for very small and very large coverage rates as well as for aspect *repetition*.

In Table 14, the mean (virtual) estimated partial process understandability, the standard deviation of the (virtual) 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 different coverage rates.

Also for these tables, we had to use a probabilistic algorithm: For each analyzed coverage rate, we randomly selected 100,000 subsets of questions. Exact values could only be computed for very small and very large coverage rates as well as for aspect *repetition*.

The different values for mean in Tables 14(a), 14(c) and 14(d) compared to Table 13 are caused by rounding errors.

Table 14 and Figure 11 support our hypothesis—aspect *repetition* having the weakest effect: For the same coverage rate, many different (virtual) 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.

For the process used in this experiment, a coverage rate of 0.25 produces less than 1% lower or upper outliers for all four aspects.

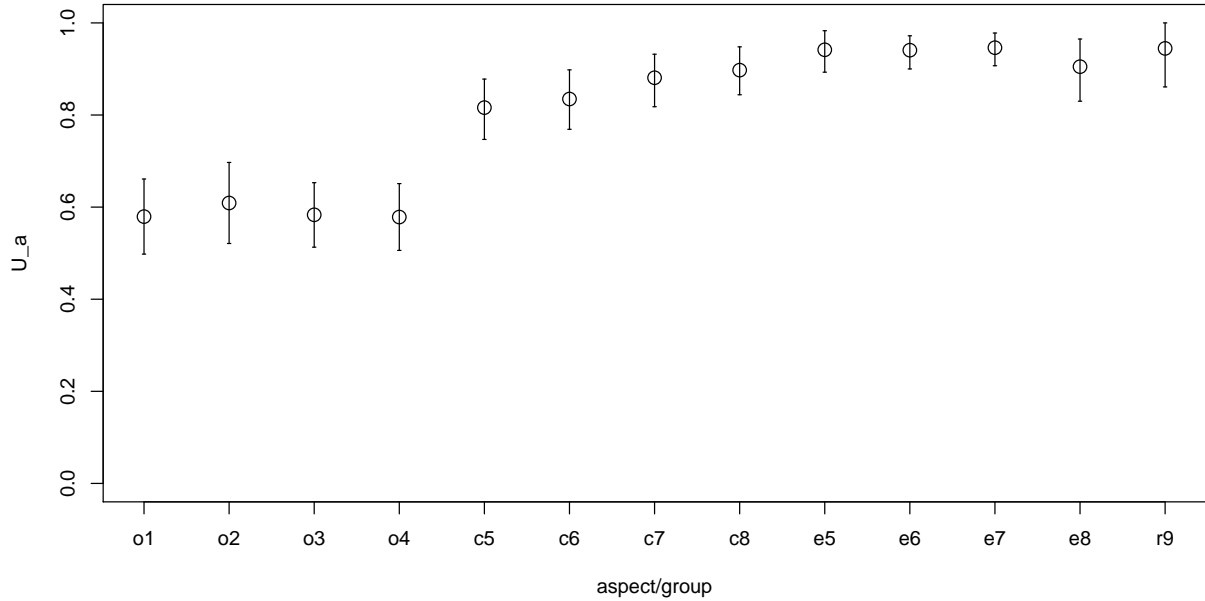
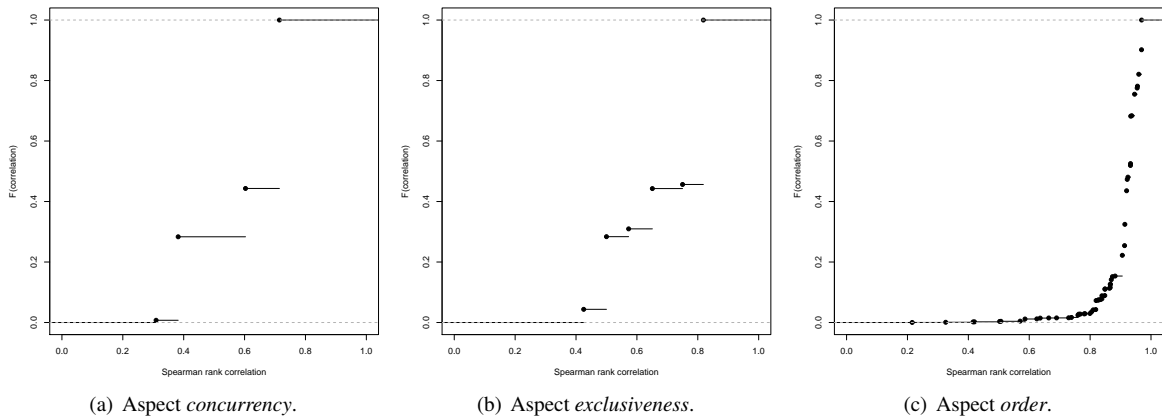


Figure 7: Estimated (partial) process understandability values and 95% confidence intervals for the 13 data sets.



(a) Aspect *concurrency*.

(b) Aspect *exclusiveness*.

(c) Aspect *order*.

Figure 8: Empirical cumulative distribution functions for Spearman rank correlations between two halves of questions.

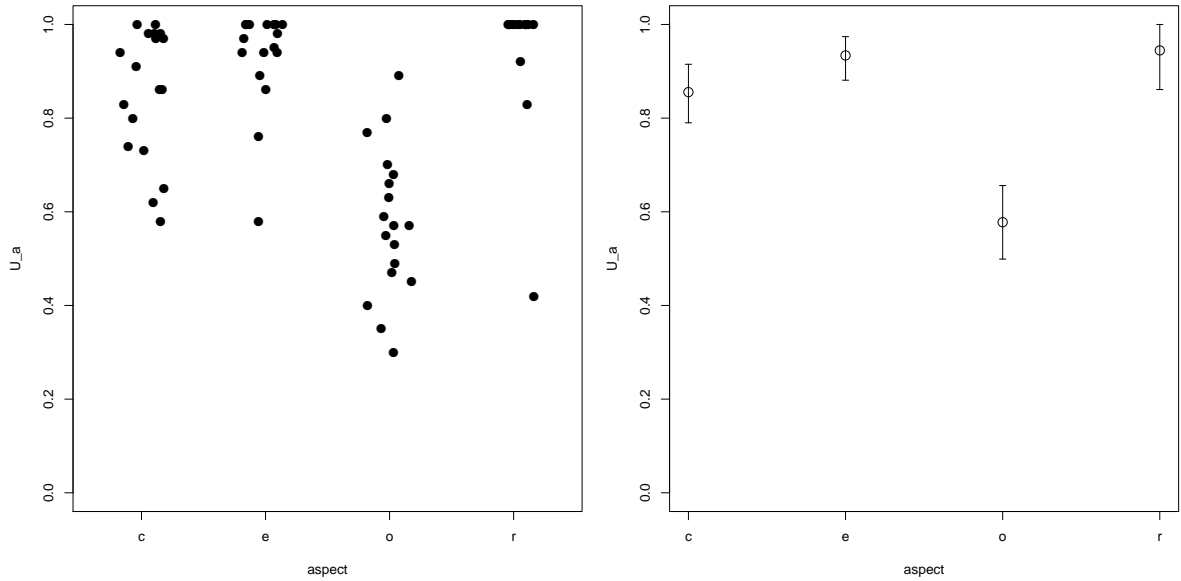
6 Conclusion and Future Work

In this paper, we gave an overview about the work on measuring structural process understandability and necessary basics about measurement and prediction systems. Based on the points of criticism expressed by Melcher and Seese about the existing measuring approaches—especially the possible lack of reliability and validity of the proposed understandability metrics—, their definitions for measuring structural process understandability and a first small experiment, we conducted a second larger experiment.

The experiment—involving 178 students from three universities—reconfirms the results of the first experiment with higher statistical significance and gives some new insights:

It supports our hypothesis that different aspects of structural process understandability are related to different levels of complexity (only *exclusiveness* and *repetition* are quite similar for our process model). Also, it turns out that using only a small part of the set of possible questions can cause values for structural process understandability differing substantially from the real value.

Therefore, all different aspects have to be included and the coverage rate of asked questions should not be too small. With respect to the process model in our experiment, a coverage rate of 0.25 resulted in less than 1% outliers (higher or lower than 95% confidence interval) for all four aspects. Finally, the asked questions should be selected at random as to minimize the risk of choosing particularly easy or difficult questions.



(a) (Virtual) personal process understandability values.

(b) (Virtual) estimated process understandability values and 95% confidence intervals.

Figure 9: Visualizations of (virtual) process understandability values.

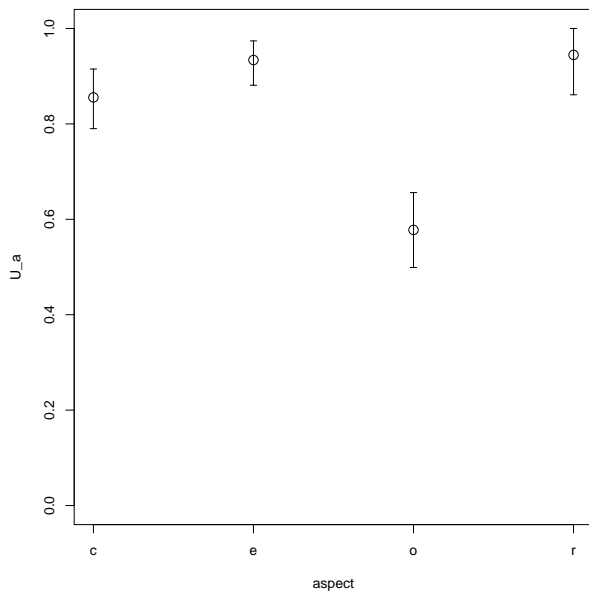


Figure 10: (Virtual) estimated process understandability values and 95% confidence intervals for the four aspects.

Our work also points to open issues that need to be addressed by future research. In our experiment, only *order* was normally distributed. This aspect also had the lowest values which is not directly intuitive. Arguably, *concurrency* and *exclusiveness* are more complicated matters than *order*. Presumably, we have to consider further characteristics of the process model in order to get an overall picture, in particular, where cer-

tain aspects matter and to what extent they can be observed in isolation when observing them in the model.

Another future issue is the selection of suitable coverage rates minimizing the measuring effort *and* the differences from the real structural process understandability value. We need to investigate whether the ideal coverage is indicated relative or absolute to the process model size and whether it depends on other (structural) process properties. It should also be examined whether there are other relevant aspects of structural understandability. Finally, once reliable and valid metrics for structural process understandability are in place, the examination of influencing factors as part of a prediction system is an important task of research.

Acknowledgment

The authors want to thank Kerstin Schmidt for her technical assistance 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] T. R. G. Green and M. Petre. Usability analysis of visual programming environments: A 'cognitive

Table 10: Correct given answers per question for the four aspects.

(a) Aspect *concurrency*.

	B	C	D	E	F	G	H	I	J	K	L
A	83%	80%	86%	95%	89%	84%	81%	90%	83%	72%	71%
B		90%	89%	76%	86%	100%	83%	76%	86%	95%	89%
C			84%	95%	90%	83%	96%	90%	85%	56%	84%
D				100%	90%	83%	92%	95%	80%	56%	84%
E					95%	80%	94%	80%	90%	70%	89%
F						84%	86%	100%	94%	88%	81%
G							90%	94%	88%	95%	95%
H								39%	80%	86%	90%
I									100%	88%	86%
J										95%	83%
K											84%

(b) Aspect *exclusiveness*.

	B	C	D	E	F	G	H	I	J	K	L
A	100%	80%	72%	100%	95%	94%	96%	90%	70%	94%	
B		92%	100%	80%	94%	100%	90%	90%	78%	88%	100%
C			100%	100%	100%	95%	100%	100%	88%	90%	100%
D				100%	96%	100%	95%	100%	84%	86%	95%
E					100%	100%	100%	95%	83%	92%	100%
F						75%	94%	92%	100%	90%	94%
G							92%	95%	100%	100%	96%
H								67%	100%	100%	96%
I									100%	95%	100%
J										92%	95%
K											80%

(c) Aspect *order*.

	A	B	C	D	E	F	G	H	I	J	K	L
A		44%	55%	86%	70%	61%	60%	81%	65%	50%	60%	81%
B	55%		67%	60%	71%	65%	61%	60%	71%	65%	61%	60%
C	62%	55%		72%	80%	48%	75%	17%	40%	38%	75%	83%
D	50%	67%	75%		83%	35%	71%	10%	28%	70%	57%	75%
E	50%	45%	62%	65%		33%	85%	24%	10%	56%	85%	62%
F	55%	50%	65%	43%	70%		6%	60%	71%	55%	89%	90%
G	62%	55%	78%	70%	48%	90%		50%	55%	43%	85%	100%
H	55%	62%	65%	44%	65%	71%	40%		83%	70%	86%	95%
I	67%	50%	52%	60%	67%	60%	67%	75%		33%	85%	86%
J	60%	56%	70%	57%	60%	33%	85%	24%	10%		94%	15%
K	67%	45%	72%	70%	48%	25%	89%	40%	24%	20%		33%
L	55%	71%	70%	61%	65%	29%	80%	33%	35%	33%	30%	

(d) Aspect *repetition*.

	A	B	C	D	E	F	G	H	I	J	K	L
A	93%	100%	93%	93%	93%	93%	87%	100%	93%	100%	87%	100%

dimensions' framework. *Journal of Visual Languages and Computing*, 7(2):131–174, 1996.

- [4] Stephen H. Kan. *Metrics and Models in Software Quality Engineering*. Addison-Wesley, 2nd edition, 2002.
- [5] 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>.
- [6] Joachim Melcher and Detlef Seese. Towards validating prediction systems for process understandability: Measuring process understandability. In Viorel Negru, Tudor Jelebelean, Dana

Petcu, and Daniela Zaharie, editors, *Proceedings of the 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2008)*, pages 564–571, 2008.

- [7] Joachim Melcher and Detlef Seese. Towards validating prediction systems for process understandability: Measuring process understandability (experimental results). Research report, Universität Karlsruhe (TH), Institut AIFB, 2008. <http://digbib.ubka.uni-karlsruhe.de/volltexte/1000009260>.
- [8] Jan Mendling. *Metrics for Process Models: Empirical Foundations of Verification, Error Prediction, and Guidelines for Correctness*, volume 6 of *Lecture Notes in Business Information Processing*. Springer, 2008.
- [9] 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.
- [10] 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.
- [11] Thomas G. Moher, David C. Mak, Brad Blumenthal, and Laura M. Leventhal. Comparing the comprehensibility of textual and graphical programs: The case of petri nets. In Curtis R. Cook, Jean C. Scholtz, and James C. Spohrer, editors, *Empirical Studies of Programmers: Fifth Workshop: Papers Presented at the Fifth Workshop on Empirical Studies of Programmers*, pages 137–161, 1993.
- [12] Michael J. Panik. *Advanced Statistics from an Elementary Point of View*. Elsevier Academic Press, 2005.
- [13] Jan Recker and Alexander Dreiling. Does it matter which process modelling language we teach or use? An experimental study on understanding process modelling languages without formal education. In *Proceedings of the 18th Australasian Conference on Information Systems (ACIS 2007)*, pages 356–366, 2007.
- [14] H. A. Reijers and J. Mendling. Modularity in process models: Review and effects. In Marlon

Table 11: Estimated (partial) process understandability values, standard deviations and 95% confidence intervals for the four aspects and nine groups.

	o1	o2	o3	o4	c5	c6	c7	c8	e5	e6	e7	e8	r9
$\bar{U}_a(p, S, Q_a)$ or $\bar{U}_r(p, S)$	0.579	0.609	0.583	0.579	0.816	0.835	0.881	0.898	0.942	0.941	0.946	0.905	0.945
s.d.	0.163	0.186	0.169	0.154	0.147	0.167	0.139	0.118	0.098	0.093	0.081	0.159	0.153
lower conf. int. bound	0.498	0.521	0.513	0.506	0.747	0.769	0.818	0.844	0.893	0.900	0.907	0.830	0.861
upper conf. int. bound	0.661	0.697	0.653	0.651	0.878	0.898	0.932	0.948	0.983	0.972	0.978	0.965	1.000

Dumas, Manfred Reichert, and Ming-Chien Shan, editors, *Business Process Management: Proceedings of the 6th International Conference BPM 2008*, volume 5240 of *Lecture Notes in Computer Science*, pages 20–35, 2008.

- [15] Kamyar Sarshar and Peter Loos. Comparing the control-flow of EPC and Petri net from the end-user perspective. In W. M. P. van der Aalst, Boualem Benatallah, Fabio Casati, and Francisco Curbera, editors, *Business Process Management: Proceedings of the 3rd International Conference BPM 2005*, volume 3649 of *Lecture Notes in Computer Science*, pages 434–439, 2005.
- [16] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4):591–611, 1965.
- [17] 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 *Lecture Notes in Computer Science*, pages 480–494, 2008.

Table 12: Construction of virtual subjects.

(a) Aspect *concurrency*.

virtual subject	subject1	subject2	subject3	subject4	$U_c^*(p, s)$
s_{01}^*	s_{082}	s_{101}	s_{126}	s_{159}	0.58
s_{02}^*	s_{096}	s_{106}	s_{140}	s_{151}	0.62
s_{03}^*	s_{089}	s_{109}	s_{124}	s_{162}	0.65
s_{04}^*	s_{081}	s_{110}	s_{128}	s_{144}	0.73
s_{05}^*	s_{083}	s_{116}	s_{142}	s_{147}	0.74
s_{06}^*	s_{080}	s_{104}	s_{127}	s_{156}	0.80
s_{07}^*	s_{086}	s_{119}	s_{139}	s_{145}	0.83
s_{08}^*	s_{087}	s_{108}	s_{129}	s_{153}	0.86
s_{09}^*	s_{094}	s_{115}	s_{130}	s_{160}	0.86
s_{10}^*	s_{090}	s_{118}	s_{136}	s_{161}	0.91
s_{11}^*	s_{092}	s_{107}	s_{123}	s_{163}	0.94
s_{12}^*	s_{093}	s_{102}	s_{125}	s_{146}	0.97
s_{13}^*	s_{095}	s_{105}	s_{131}	s_{148}	0.97
s_{14}^*	s_{085}	s_{112}	s_{132}	s_{149}	0.98
s_{15}^*	s_{088}	s_{113}	s_{134}	s_{150}	0.98
s_{16}^*	s_{091}	s_{114}	s_{135}	s_{152}	0.98
s_{17}^*	s_{084}	s_{120}	s_{141}	s_{154}	1.00
s_{18}^*	s_{097}	s_{121}	s_{143}	s_{155}	1.00

(b) Aspect *exclusiveness*.

virtual subject	subject1	subject2	subject3	subject4	$U_e^*(p, s)$
s_{19}^*	s_{091}	s_{109}	s_{132}	s_{159}	0.58
s_{20}^*	s_{080}	s_{114}	s_{125}	s_{147}	0.76
s_{21}^*	s_{085}	s_{106}	s_{135}	s_{151}	0.86
s_{22}^*	s_{089}	s_{110}	s_{140}	s_{149}	0.89
s_{23}^*	s_{083}	s_{115}	s_{123}	s_{144}	0.94
s_{24}^*	s_{084}	s_{118}	s_{124}	s_{154}	0.94
s_{25}^*	s_{093}	s_{119}	s_{128}	s_{160}	0.94
s_{26}^*	s_{097}	s_{120}	s_{129}	s_{145}	0.95
s_{27}^*	s_{081}	s_{121}	s_{131}	s_{146}	0.97
s_{28}^*	s_{082}	s_{101}	s_{136}	s_{148}	0.98
s_{29}^*	s_{086}	s_{102}	s_{126}	s_{150}	1.00
s_{30}^*	s_{087}	s_{104}	s_{127}	s_{152}	1.00
s_{31}^*	s_{088}	s_{105}	s_{130}	s_{153}	1.00
s_{32}^*	s_{090}	s_{107}	s_{134}	s_{155}	1.00
s_{33}^*	s_{092}	s_{108}	s_{139}	s_{156}	1.00
s_{34}^*	s_{094}	s_{112}	s_{141}	s_{161}	1.00
s_{35}^*	s_{095}	s_{113}	s_{142}	s_{162}	1.00
s_{36}^*	s_{096}	s_{116}	s_{143}	s_{163}	1.00

(c) Aspect *order*.

virtual subject	subject1	subject2	subject3	subject4	$U_o^*(p, s)$
s_{37}^*	s_{017}	s_{029}	s_{046}	s_{073}	0.30
s_{38}^*	s_{007}	s_{023}	s_{047}	s_{061}	0.35
s_{39}^*	s_{010}	s_{025}	s_{055}	s_{065}	0.40
s_{40}^*	s_{002}	s_{020}	s_{053}	s_{063}	0.45
s_{41}^*	s_{012}	s_{036}	s_{039}	s_{079}	0.47
s_{42}^*	s_{014}	s_{032}	s_{041}	s_{064}	0.49
s_{43}^*	s_{008}	s_{034}	s_{043}	s_{074}	0.53
s_{44}^*	s_{011}	s_{019}	s_{054}	s_{075}	0.55
s_{45}^*	s_{006}	s_{028}	s_{057}	s_{068}	0.57
s_{46}^*	s_{013}	s_{030}	s_{058}	s_{070}	0.57
s_{47}^*	s_{004}	s_{031}	s_{048}	s_{062}	0.59
s_{48}^*	s_{005}	s_{024}	s_{056}	s_{066}	0.63
s_{49}^*	s_{003}	s_{033}	s_{045}	s_{077}	0.66
s_{50}^*	s_{015}	s_{021}	s_{052}	s_{071}	0.68
s_{51}^*	s_{016}	s_{038}	s_{051}	s_{072}	0.70
s_{52}^*	s_{001}	s_{035}	s_{042}	s_{078}	0.77
s_{53}^*	s_{009}	s_{026}	s_{050}	s_{060}	0.80
s_{54}^*	s_{018}	s_{022}	s_{059}	s_{076}	0.89

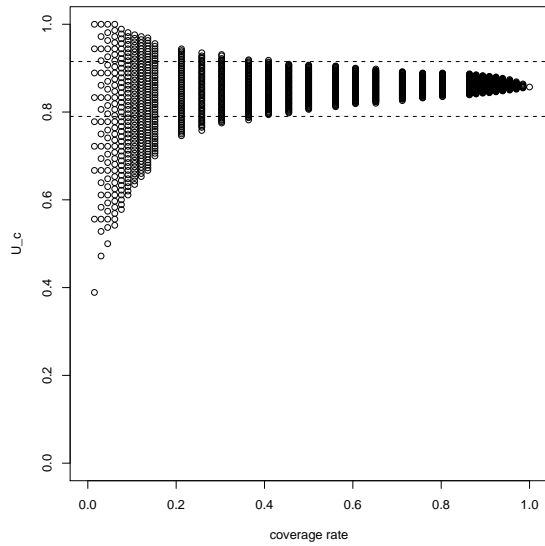
Table 13: (Virtual) estimated process understandability values, standard deviations and 95% confidence intervals for the four aspects.

	concurrency	exclusiveness	order	repetition
$\widehat{U}_a^*(p, S)$ or $\widehat{U}_r^*(p, S)$	0.856	0.934	0.578	0.945
s.d.	0.140	0.109	0.157	0.153
lower conf. int. bound	0.790	0.881	0.499	0.861
upper conf. int. bound	0.915	0.974	0.656	1.000

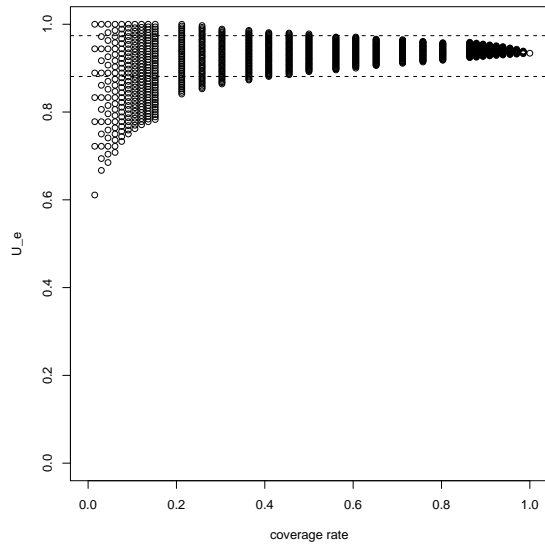
Table 14: Data about (virtual) estimated partial process understandability values for the four aspects.

(a) Aspect <i>concurrency</i> .						(b) Aspect <i>exclusiveness</i> .					
# questions	cov. rate	mean	s. d.	rate lower	rate higher	# questions	cov. rate	mean	s. d.	rate lower	rate higher
1	0.02	0.857	0.105	16.7%	27.3%	1	0.02	0.934	0.082	16.7%	43.9%
2	0.03	0.857	0.073	12.9%	25.1%	2	0.03	0.934	0.057	14.7%	18.9%
3	0.05	0.857	0.059	12.2%	11.4%	3	0.05	0.934	0.046	12.7%	22.2%
4	0.06	0.857	0.051	10.0%	11.4%	4	0.06	0.934	0.040	10.7%	11.4%
5	0.08	0.857	0.045	10.2%	5.4%	5	0.08	0.934	0.035	8.9%	13.1%
7	0.11	0.857	0.037	5.3%	2.6%	7	0.11	0.934	0.029	5.9%	8.1%
10	0.15	0.857	0.031	2.8%	1.2%	10	0.15	0.934	0.024	2.0%	2.8%
14	0.21	0.857	0.025	0.7%	0.2%	14	0.21	0.934	0.019	0.7%	1.0%
17	0.26	0.857	0.022	0.1%	0.1%	17	0.26	0.934	0.017	0.1%	0.3%
20	0.30	0.857	0.019	0.0%	0.0%	20	0.30	0.934	0.015	0.0%	0.2%
24	0.36	0.857	0.017	0.0%	0.0%	24	0.36	0.934	0.013	0.0%	0.0%
27	0.41	0.857	0.015	0.0%	0.0%	27	0.41	0.934	0.012	0.0%	0.0%
30	0.45	0.857	0.014	0.0%	0.0%	30	0.45	0.934	0.011	0.0%	0.0%
33	0.50	0.857	0.013	0.0%	0.0%	33	0.50	0.934	0.010	0.0%	0.0%
37	0.56	0.857	0.011	0.0%	0.0%	37	0.56	0.934	0.009	0.0%	0.0%
40	0.61	0.857	0.010	0.0%	0.0%	40	0.61	0.934	0.008	0.0%	0.0%
43	0.65	0.857	0.009	0.0%	0.0%	43	0.65	0.934	0.007	0.0%	0.0%
47	0.71	0.857	0.008	0.0%	0.0%	47	0.71	0.934	0.006	0.0%	0.0%
50	0.76	0.857	0.007	0.0%	0.0%	50	0.76	0.934	0.006	0.0%	0.0%
53	0.80	0.857	0.006	0.0%	0.0%	53	0.80	0.934	0.005	0.0%	0.0%
57	0.86	0.857	0.005	0.0%	0.0%	57	0.86	0.934	0.004	0.0%	0.0%
60	0.91	0.857	0.004	0.0%	0.0%	60	0.91	0.934	0.003	0.0%	0.0%
62	0.94	0.857	0.003	0.0%	0.0%	62	0.94	0.934	0.003	0.0%	0.0%
63	0.95	0.857	0.003	0.0%	0.0%	63	0.95	0.934	0.002	0.0%	0.0%
64	0.97	0.857	0.002	0.0%	0.0%	64	0.97	0.934	0.002	0.0%	0.0%
65	0.98	0.857	0.002	0.0%	0.0%	65	0.98	0.934	0.001	0.0%	0.0%
66	1.00	0.857	—	0.0%	0.0%	66	1.00	0.934	—	0.0%	0.0%

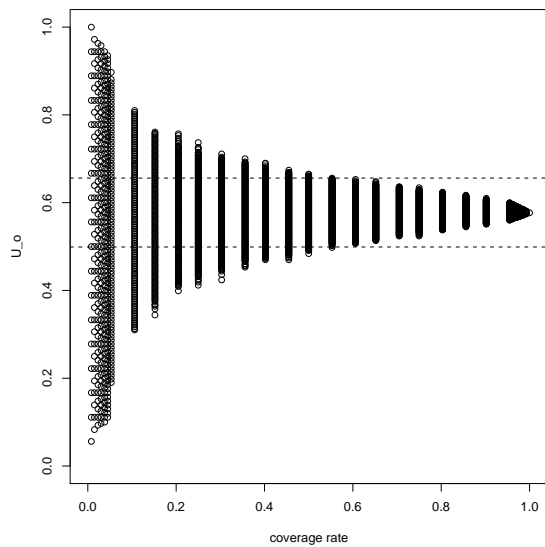
(c) Aspect <i>order</i> .						(d) Aspect <i>repetition</i> .					
# questions	cov. rate	mean	s. d.	rate lower	rate higher	# questions	cov. rate	mean	s. d.	rate lower	rate higher
1	0.01	0.577	0.210	28.8%	47.0%	1	0.08	0.944	0.048	0.0%	0.0%
2	0.02	0.577	0.147	26.2%	33.6%	2	0.17	0.944	0.031	0.0%	0.0%
3	0.02	0.577	0.120	23.3%	26.9%	3	0.25	0.944	0.024	0.0%	0.0%
4	0.03	0.577	0.103	20.7%	22.2%	4	0.33	0.944	0.020	0.0%	0.0%
7	0.05	0.577	0.077	14.9%	16.1%	5	0.42	0.944	0.016	0.0%	0.0%
14	0.11	0.577	0.053	7.2%	6.4%	6	0.50	0.944	0.014	0.0%	0.0%
20	0.15	0.577	0.043	3.7%	3.0%	7	0.58	0.944	0.012	0.0%	0.0%
27	0.20	0.577	0.036	1.6%	1.3%	8	0.67	0.944	0.010	0.0%	0.0%
33	0.25	0.578	0.032	0.7%	0.5%	9	0.75	0.944	0.008	0.0%	0.0%
40	0.30	0.578	0.028	0.3%	0.2%	10	0.83	0.944	0.006	0.0%	0.0%
47	0.36	0.578	0.025	0.1%	0.0%	11	0.92	0.944	0.004	0.0%	0.0%
53	0.40	0.577	0.022	0.0%	0.0%	12	1.00	0.944	—	0.0%	0.0%
60	0.45	0.577	0.020	0.0%	0.0%						
66	0.50	0.578	0.018	0.0%	0.0%						
73	0.55	0.577	0.016	0.0%	0.0%						
80	0.61	0.577	0.015	0.0%	0.0%						
86	0.65	0.577	0.013	0.0%	0.0%						
93	0.70	0.577	0.012	0.0%	0.0%						
99	0.75	0.577	0.011	0.0%	0.0%						
106	0.80	0.577	0.009	0.0%	0.0%						
113	0.86	0.577	0.007	0.0%	0.0%						
119	0.90	0.577	0.006	0.0%	0.0%						
126	0.95	0.577	0.004	0.0%	0.0%						
128	0.97	0.577	0.003	0.0%	0.0%						
129	0.98	0.577	0.003	0.0%	0.0%						
130	0.98	0.577	0.002	0.0%	0.0%						
131	0.99	0.577	0.002	0.0%	0.0%						
132	1.00	0.577	—	0.0%	0.0%						



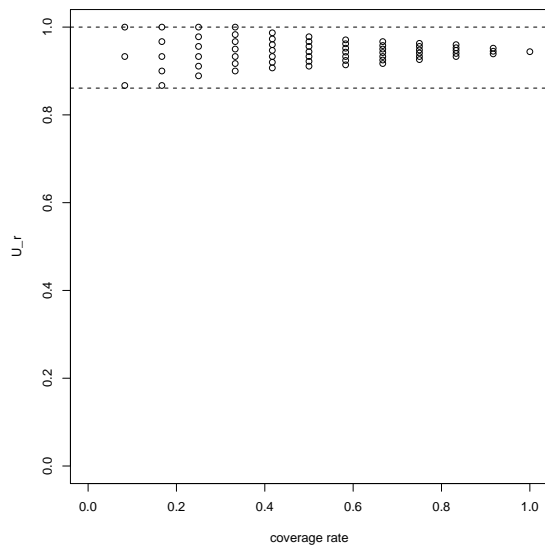
(a) Aspect *concurrency*.



(b) Aspect *exclusiveness*.



(c) Aspect *order*.



(d) Aspect *repetition*.

Figure 11: (Virtual) estimated partial process understandability values of the four aspects depending on coverage rate.