Eye-Tracking the Factors of Process Model Comprehension Tasks

Razvan Petrusel and Jan Mendling

Faculty of Economics and Business Administration, Babes-Bolyai University,
Teodor Mihali str. 58-60, 400591 Cluj-Napoca, Romania
razvan.petrusel@econ.ubbcluj.ro
Wirtschaftsuniversität Wien, Augasse 2-6, 1090 Wien, Austria
jan.mendling@wu.ac.at

Abstract. Understanding business process models has been previously related to various factors. Those factors were determined using statistical approaches either on model repositories or on experiments based on comprehension questions. We noticed that, when asking comprehension questions on a process model, usually the expert explores only a part of the entire model to provide the answer. This paper formalizes this observation under the notion of Relevant Region. We conduct an experiment using eye-tracking to prove that the Relevant Region is indeed correlated to the answer given to the comprehension question. We also give evidence that it is possible to predict whether the correct answer will be given to a comprehension question, knowing the number and the time spent fixating Relevant Region elements. This paper sets the foundations for future improvements on model comprehension research and practice.

Keywords: process model comprehension factors, process model relevant region, process model eye tracking experiment.

1 Introduction

Although business process modeling has become widely adopted and intensively researched in the last decade, we still know quite little about the concrete act of sensemaking while a human inspects a model. At least, prior research has identified various factors that have an influence on how well a process model is understood. These factors mainly relate to the representation of the model, for instance its size, its complexity and its notation, and to characteristics of the person reading the model, including modeling expertise or familiarity with a particular modeling language [1], [2], [3], [4]. Most of these factors can be traced back to theories such as cognitive load theory.

What is striking in this context is the fact that the comprehension performance of a particular person in interpreting a specific model can be quite diverging. While it has been demonstrated that comprehension tasks vary in their degree of difficulty [5], insights into the set of potential task-related factors is rather limited and partially inconclusive. Specifically, a distinction upon the types of behavior (sequence, concurrency, exclusiveness) has not worked well to separate easy from difficult

C. Salinesi, M.C. Norrie, and O. Pastor (Eds.): CAiSE 2013, LNBIP 7908, pp. 224–239, 2013.

comprehension tasks [6]. Beyond that, such a distinction does not help in explaining the striking importance of the degree of structuredness for comprehension. The notion of structuredness measures if a model is built using nested blocks of matching join and split connectors [7], [8].

In this paper we address the gap of research on the factors that influence the comprehension tasks. We approach this topic from the perspective of both the process model and the comprehension tasks together, which provides us with a basis for defining the notion of a so-called Relevant Region and its components, the Relevant Model Elements. In order to evaluate the significance of this notion, we use an eye-tracking experiment with expert process modelers. Our results confirm the relevance of this notion, which has implications for future model comprehension experiments and for improving model comprehension in practice.

The paper is structured as follows. Section 2 discusses the background of our research. We summarize findings in this area and standard ways of measuring comprehension performance. This provides us with the basis to define the notion of a Relevant Region. Section 3 presents our research design. We formalize our expectations in terms of four hypotheses. Then, we present the experimental design for investigating these hypotheses, and the experimental setup. Section 4 presents the results of the experiment. We summarize the demographics of the participants. Furthermore, we utilize correlational analysis to inspect the hypotheses, and logistic regression to predict the probability of a correct answer based on the Relevant Region metrics. Section 5 discusses our findings in the light of related work, before Section 6 concludes.

2 Background

In this section, we discuss the background of our research. First, we revisit the foundations of process model comprehension performance. Then, we describe novel directions for the definition of comprehension task.

2.1 Process Model Comprehension Performance

Model comprehension is an important facet that is closely associated with a more general notion of model quality. According to semiotic theory and its adoptions to model quality, a reader of a model has to understand the syntax and the semantics of a model correctly in order to be able to draw correct pragmatic conclusions from it [9]. Comprehension of a model cannot be directly observed. Therefore, comprehension performance is typically approached by providing tasks to a model reader that can only be solved correctly when the model is well understood. The range of potential tasks types includes cloze tests, problem solving tasks, or speed of answering questions [10]. More specifically, in the area of process model comprehension, interpretation tasks relating to the formal behavior are typically used to operationalize comprehension [4], [11]. Such interpretation tasks can be constructed by presenting a process model to a model reader and asking how a specific pair of activities is related from a behavioral point of view (being concurrent, exclusive or ordered). The correct solutions can be automatically checked based on the formal semantics of the model [12].

Fig. 1 shows the example of a process model which was part of the BPMN Selftest (more details are available at http://granturi.ubbcluj.ro/decision_mining/docs/BPMN-Selftest-Material.pdf). People could participate in this self-test by running through a series of process model comprehension questions on a website. In relation to the model shown in the figure, it is interesting to note that user characteristics and model characteristics alone are hardly able to explain the comprehension performance. For instance, the comprehension task can *Z and AA be executed in the same case (yes/no)* was answered correctly by 65.6% of 430 participants, while the same participants had 72.2% correct answers for the task *After O has been executed, and the default path is taken at the next gateway, then Z must always be executed (yes/no).* Since we randomly sampled the questions, we can rule out fatigue as a distorting factor. This sampling was organized in such a way that participants never got tasks on the same model directly after one another to avoid memorizing bias. Furthermore, the second question is 13 words longer than the first one, which should imply a higher burden in terms of cognitive load. Still, on the aggregate level it was easier for participants to give a correct answer to the second question. Next, we define metrics that might serve as factors of model comprehension.

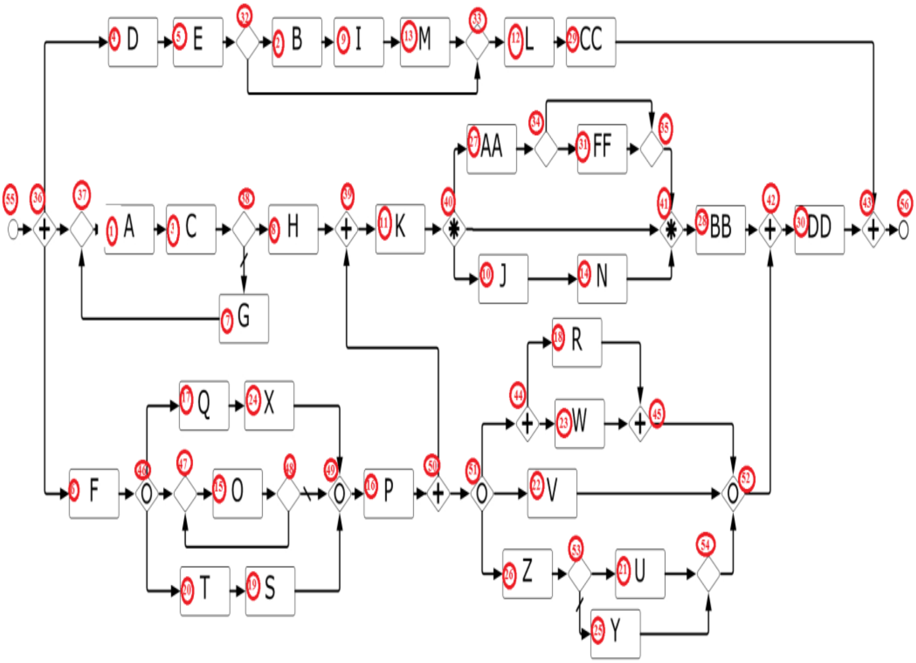


Fig. 1. Example model from the BPMN Selftest annotated with element numbers

2.2 Model Comprehension and the Notion of a Relevant Region

The idea of defining new metrics in relation to model comprehension relies on the observation that not the whole model has to be studied for providing a correct solution to a comprehension task. If we consider the second task referring to O and Z, we find that we can easily find a path from O to Z with only four gateways and one activity in between. In contrast to that, the relationship between Z and AA is much more difficult to assess. To this end, we have to inspect the model in a backward mode from the two activities up to the gateway from which both paths to Z and AA originate. Here, this is the split gateway (labeled 36) in the very left part of the model. If nodes 36, 39 and 50 were exclusive choice gateways (XOR-splits), the answer would be *no*. As we observe an AND-split, the correct answer is *yes*. The challenge in finding this solution is that a considerably larger share of the model has to be inspected as for the task on O and Z. From the AND-split to AA, we have to traverse the model via A, C, H, and K; from the split to Z, we pass at least F, O, and P. This observation could explain the better results for the second question introduces in the previous sub-section. The goal of this paper is to formalize this observation, and then investigate if it can be generalized.

In line with this observation, we formalize a notion of Relevant Region as a potential factor of model comprehension. This formalization is based on the definition of a process model, its notion of path, and the notion of a dominator (cf. [13]). A *process model* is defined as a tuple $PM = (N, F, l)$ with N being the set of nodes and the arcs defined as $F \subseteq N \times N$. The set of nodes is partitioned into mutually disjoint sets as $N = \{s, e\} \cup A \cup G$ referring to start and end events, activities and gateways. The function $l: G \rightarrow \{AND, OR, XOR\}$ maps gateways to corresponding label types AND, OR, and XOR. A *path* $n_1 \rightsquigarrow n_k$ is a non-empty sequence of nodes $n_1, \dots, n_k$ such that $(n_1, n_2), \dots, (n_{k-1}, n_k) \in F$. For two nodes x, y we define x as a *dominator* of y, if and only if for all paths from the start event $s \rightsquigarrow y$ it holds that $x \in (s \rightsquigarrow y)$. The *dominating node* $dom(x, y) = z$ is that node that is both a dominator to x and y, and which has no other dominators on its paths to x and y (cf. the notion of start join in [14]). In a process structure tree decomposition of a process model, each single-entry node of a fragment is a dominating node for all pairs of nodes within this fragment.

Based on these notions, we can establish the definition of a *Relevant Region* $RR(a, b, PM) \subseteq N$ such that

$$RR(a, b, PM) = \{n \in N \mid n \in dom(a, b) \rightsquigarrow a \vee dom(a, b) \rightsquigarrow b\}.$$

Consider the example of Fig. 1 where each node is numbered for easy reference. Given a comprehension task as *Can R and W show up in the same case? (yes/no)*, we observe that R and W share a dominating node $dom(R, W) = n_{44}$, which is the AND-gateway directly before them. Accordingly, $RR(R, W, PM) = \{R, W, n_{44}\}$. By inspecting the elements and connections in this area, we find the correct answer is *yes*.

The significance of this notion of a *Relevant Region* can be investigated from two angles. First, with a focus on process model comprehension results, it can be checked whether a variation in the Relevant Region of a task is associated with a variation in comprehension performance. Second, it can be approached by mapping the comprehension process onto the process model. The latter can be facilitated using eye-tracking. Indeed, eye-tracking has been recently proposed as a technique for investigating the cognitive process of model comprehension on a more fine-granular level as compared to existing approaches which consider task results only [15], [16].

In order to correctly understand a model, an individual has to inspect the elements depicted in the model. Looking at model elements is highly correlated with the individual's thinking process [17]. Eye-tracking equipment helps to create a record of the elements the subject's eyes fixated upon. Other interesting data extracted using this method is the fixation sequence and fixation times of the different elements. Fixation means that a person's eyes are aimed at some object, therefore he investigates it. Fixation sequence is the order of the items a person looks at. Fixation time is the period of time over which the subject's eyes are directed at the object. In our eye-tracking experiment, subjects look at process models. Therefore, a fixation occurs when the subject looks at the model node for a period of time over a certain threshold that will allow his brain to capture the meaning of the visual stimulus [18], [19], [20]. The eye-tracking software calculates fixation time as the length of time the eye velocity was below both the saccade velocity criterion and the drift distance criterion. Saccades are fast rotations of the eyes that occur several times each second and are commanded automatically by the brain (without getting awareness) [20]. Saccades show up when the subject's attention shifts from one point on the screen to another. Using recorded fixation sequences, we can define the set of elements that a subject has looked at, as the notion of a Scan Path $SP \subseteq N$. This Scan Path, SP, is the set of nodes of a process model that get a fixation from the subject's eyes.

Based on both the notion of a Scan Path SP and concepts from information retrieval, we can discuss the appropriateness of the Relevant Regions RR concept. The area of information retrieval focuses on evaluating techniques that provide a number of results from a given search space. This perspective can be directly related to our research problem. The standard notions of precision, recall and f-measure can be adapted accordingly [21]. Simply put, precision is the percentage of relevant items from all retrieved items while recall is the percentage of relevant retrieved items from all relevant items. The f-measure is the harmonic mean of precision and recall. Given a process model and a subject that fixates some of the model elements in order to answer a comprehension question, we can identify the corresponding Relevant Region RR. Together with the actually observed Scan Path, we get Scan Path Precision (SPP), Scan Path Recall SPR, and Scan Path f-measure (SPF):

$$SPP = \frac{SP \cap RR}{SP}, SPR = \frac{SP \cap RR}{RR}, SPF = 2 * \frac{SPP * SPR}{SPP + SPR}$$

Based on these concepts, we have defined the foundations to empirically test the significance of the notion of a Relevant Region.

3 Research Design

This section introduces the experimental setup. We first define the researched hypotheses, and then give an overview of the methodology employed to validate them. At the core of our approach lays the notion of Relevant Region introduced in the previous section and the experiments involving tracking the subject's fixations on (looks at) the elements of the model (eye-tracking).

H1: The Relevant Region elements are fixated a longer time than other model elements by the subjects that provided the correct answer to the comprehension question;
H2: More elements of the Relevant Region are fixated than other model elements by the subjects that provided the correct answer to the comprehension question;
H3: The higher the percentage of time spent fixating the Relevant Region elements, the more likely is a correct answer;
H4: The higher the share of Relevant Region elements a person fixates (scan-path recall and/or f-measure), the more likely is a correct answer.

In order to prove our hypotheses we follow different approaches. First, we gather experimental eye-tracking data from live experiments. Then, for H1 and H2 we do a statistical correlation analysis of the data. For H3 and H4, we model a logistic regression for estimating the probability of giving the (binary) correct answer. As a follow-up to H3 and H4 we try to discover a model that will predict the probability of providing a correct answer to a comprehension question.

3.1 Participants

Previous research showed expertise plays an important role in process model comprehension [4]. Therefore, we decided to use only experts as subjects for our experiments. There were several experimental sessions stretched between August 2012 and November 2012 with a total of 26 process model experts recruited both from academia and industry. Academia experts included in those sessions were selected from the Babes-Bolyai University in Cluj-Napoca (UBB), the Wirtschaftsuniversität Wien (WU) and from the Technical University in Eindhoven (TUE). Sessions including industry experts were organized during the 4th International Workshop on BPMN held in Vienna. This selection of subjects covers multiple backgrounds: subjects from UBB and from industry have no focus on a specific process modeling notation, subjects from WU are more familiar with BPMN while subjects from TUE use mainly Petri Nets. Given the expertise level, each subject was (highly) qualified to answer the comprehension questions. To evaluate the level of expertise, each subject was asked to fill-in a self-evaluation questionnaire as the one used in [4].

The evaluated variables are: Models read in the last year (MR) which ranges from 0 to maximum 100, Models created in the last year (MC), Familiarity with understanding and using BPMN (FAM) which ranges from 1 (very much) to 7 (none), Modeling years (MY) and the number of months since using BPMN (MBPMN). The synthetic data giving the lowest value/highest value/mean/standard_deviation is introduced in Table 1.

As can be seen from Table 1, the total level of expertise is high given that, the average number of months the subjects used BPMN is 36, they are familiar with the notation (2.6 on a scale from 1 to 7) and have read an average of 61 process models.

Table 1. Subject expertise level

Variable	*Cluj-Napoca*	*Vienna*	*Eindhoven*	*Total*
Size	4	15	7	26
MR	20/100/49/36.6	5/120/54/40	30/100/82.9/29.8	5/120/60.9/38.2
MC	5/50/22.5/20.2	2/100/23.2/24.9	20/100/54.2/33.1	2/100/31.5/29.3
FAM	1/5/2.7/2.9	1/5/2.2/1.2	1/5/3.3/1.7	1/5/2.6/1.5
MY	1/5/2.5/3.2	2.5/10/6.6/2.6	4/8/5.7/1.3	1/10/5.8/2.6
MBPMN	12/60/30/21.3	5/72/37.4/18.1	3/60/34/24.6	0/72/36.3/20.7

3.2 Measured Variables

The independent variables measured based on the eye-tracking output data are divided according to the two investigated dimensions:

- the number of elements in the Relevant Region fixated by the subject. To correlate the number of RR elements fixated with the total number of elements fixated we calculate scan-path precision (SPP), scan-path recall (SPR), scan-path F-measure (SPF), and scan-path F2-measure (SPF2);
- the fraction of the model investigation time spent fixating each Relevant Region element (Time In Region – TIR). This variable is calculated as the time spent fixating one model element in RR over the total time spent fixating all model elements.

The dependent variable is Outcome. It is a binary variable that shows if the subject provided the correct (1) or the incorrect answer (0) to the comprehension question.

3.3 Experiment Implementation Details

The experiment was performed in seven steps as follows:

a) *Hardware set-up*. For experimenting we used a fixed-head eye-tracking system produced by Arrington Research (http://www.arringtonresearch.com/headfixed.html). Some pictures taken during the experiments that show the hardware setup are available at: http://granturi.ubbcluj.ro/decision_mining/experimente-en.html;

b) *Calibration*. This is an essential step that influences data accuracy. Calibration means mapping eye vectors (left and right) to a position on the screen. For the experiments, we calibrated a number of 42 points to balance between high fidelity (more calibration points is better) and time (more calibration points require a longer calibration period in which the subject might become tired and/or bored).

c) *Calibration confirmation*. This step give assurance over the calibration quality;

d) *Show BPMN model and ask comprehension question*. Recording eye movements starts when the model is displayed on screen and a comprehension question is asked;

e) *Record question answer*. The subject says out loud the answer to the comprehension question. All answers are Boolean (True or False). The eye movements recording stops once the answer is given;

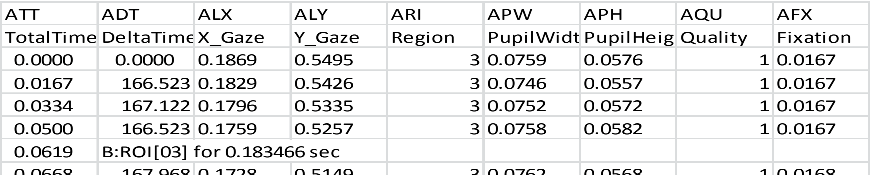
f) *Slip correction*. After each question, a quick re-calibration (slip-correction) is performed. The basic idea is to compensate the subject's minor head movements (e.g. while speaking out loud the answer).

g) *Skip to the next question*. Typically, we repeated steps c) through f) for each of the six comprehension questions. In rare cases (under 10%), once a full calibration succeeded, there was a need to also repeat step b) later in the experiment.

The experiment used a set of 5 models. Each has a structured and an unstructured version (the material can be downloaded from http://granturi.ubbcluj.ro/decision _mining/experimente.html). We asked a set of 6 questions covering those models (the two questions using the same model were placed first and last). All 26 subjects were given the basic treatment (2 comprehension questions from structured models and 4 from the unstructured ones). Three subjects were also given, in a different day, the alternate treatment (i.e. were asked the same questions but on the 'other' model). None of the alternate treatment subjects reported a memory effect.

3.4 Experimental Data

To better understand the data outputted by the eye-tracking system we will use first a small running example. The output of the experiment is a data file as shown in Fig. 2 that stores separate data for the left eye (A) and for the right eye (B). Further data includes the timestamp (ATT), the elapsed time between eye movements (ADT), the X and Y coordinates of the pupil (ALX and ALY), which region of interest the eye coordinates are placed in (ARI), pupil width (APW), height (APH), the quality of the pupil detection (AQU) and how much time the eye didn't move, in seconds, (AFX). The log also records events like eyes fixating a ROI, Fixations, Drifts and Saccades, individually for each eye (A or B).



ATT	ADT	ALX	ALY	ARI	APW	APH	AQU	AFX
TotalTime	DeltaTime	X_Gaze	Y_Gaze	Region	PupilWidt	PupilHeig	Quality	Fixation
0.0000	0.0000	0.1869	0.5495	3	0.0759	0.0576	1	0.0167
0.0167	166.523	0.1829	0.5426	3	0.0746	0.0557	1	0.0167
0.0334	167.122	0.1796	0.5335	3	0.0752	0.0572	1	0.0167
0.0500	166.523	0.1759	0.5257	3	0.0758	0.0582	1	0.0167
0.0619	B:ROI[03] for 0.183466 sec							
0.0668	167.968	0.1728	0.5149	3	0.0762	0.0568	1	0.0168

Fig. 2. Partial eye-tracking data log for a run of the experiment

The data stored in the file introduced in Fig. 2 enables the post-hoc replay over the stimulus (model). The replay of the partial trace introduced in Fig. 2 is presented in Fig. 3. It is explicitly depicting the behavior of an expert while answering the question "Can R and W be executed for the same case?".

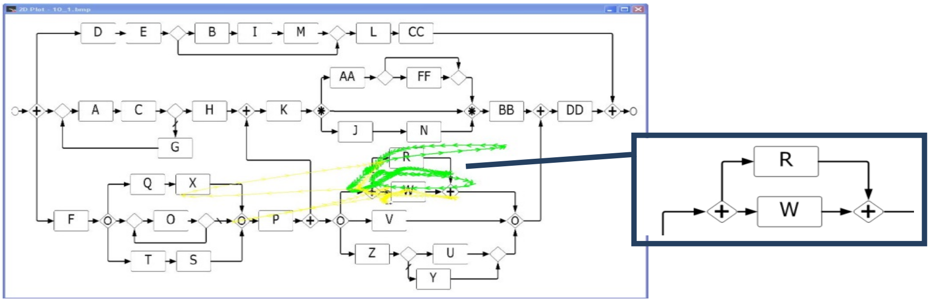
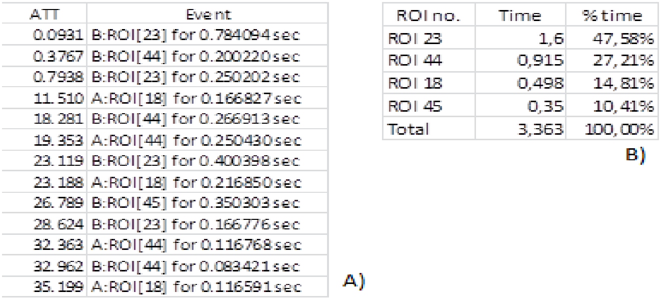


Fig. 3. Post-hoc Replay Result

In order to ease the analysis we make some assumptions. First, the sequence of fixations is not important. Second, we abstract from the count of the number of fixations for one element and use just the total time a ROI is fixated. Instead, we keep the aggregate value (e.g. that the user fixated ROI 23 for 1.58 seconds). In this way, we convert the log in Fig. 4A to the synthetic data in Fig. 4B.



ATT	Event
0.0931	B:ROI[23] for 0.784094 sec
0.3767	B:ROI[44] for 0.200220 sec
0.7938	B:ROI[23] for 0.250202 sec
11.510	A:ROI[18] for 0.166827 sec
18.281	B:ROI[44] for 0.266913 sec
19.353	A:ROI[44] for 0.250430 sec
23.119	B:ROI[23] for 0.400398 sec
23.188	A:ROI[18] for 0.216850 sec
26.789	B:ROI[45] for 0.350303 sec
28.624	B:ROI[23] for 0.166776 sec
32.363	A:ROI[44] for 0.116768 sec
32.962	B:ROI[44] for 0.083421 sec
35.199	A:ROI[18] for 0.116591 sec

A)

ROI no.	Time	%time
ROI 23	1,6	47,58%
ROI 44	0,915	27,21%
ROI 18	0,498	14,81%
ROI 45	0,35	10,41%
Total	3,363	100,00%

B)

Fig. 4. A) Filtered log showing sequence and duration of fixations in the Regions of Interest; B) Synthesis eye-tracking data

There are some risks that threat the validity of results and may limit our conclusions:

- *eye-tracking hardware and software imprecision*. It is inherent to any device and is due to the hardware limitations (e.g. video recording speed) and/or to the algorithms used to calculate the position of gaze. The threat is that there could be slight differences between the exact coordinate fixated by the subject and the one recorded in the log. To mitigate this risk we used models with enough distance between elements and we defined ROIs slightly larger than the actual model element.

- *de-calibration during experiment*. This is a serious risk which leads to the rejection of the entire observation. We used fixed-head eye-tracking system (i.e. the user's head is fixed in the chin and nose areas) but still, head movements will cause de-calibration (i.e. the user looks at one element but the software logs another or a coordinate outside the screen area). To mitigate this risk we did a post-hoc visual examination of each eye-movie and rejected those that obviously been de-calibrated. The percentage of rejected observations was 10.35% of all traces (18 out of 174).

- *personal biological features*. For most humans, one eye is dominant focusing first on the model element while the other eye lags behind. Therefore, the eyes move over different lines in the scan-path visualization (see Fig. 3). Also, there are cases in which one eye focuses on one model element while the other focuses on an adjacent one for a brief moment. To mitigate this risk we recorded both eyes independently. Then, we calculated the subject's scan-path as the union of ROIs visited by both eyes.

4 Results

One way of examining the experimental data is to strictly evaluate the percentage of correct answers to each comprehension question. The share of correct answers is in the same range as for the prior experiments with the same questions without eye-tracking [4]. To rule out structuredness of models as a factor influencing our results, we investigated an evenly distributed number of structured (e.g. model no 30, 50) and unstructured ones (e.g. 19, 29, 39). Some of the results are introduced in Table 2. As one can note, for a question we recorded a large number of incorrect answers.

Table 2. Answer correctness to comprehension questions

Model_question no	*19_0*	*19_6*	*29_5*	*30_3*	*39_6*	*50_1*
Correct (no.)	18	18	17	6	16	14
Incorrect (no.)	5	6	2	10	0	9
Correct (%)	78.26%	66.67%	89.47%	37.5%	100%	60.87%

Table 3. Sample from the eye-tracking aggregated data

Subject	*1*	*2*	*3*	*15*	*25*
Question code	10_0	10_0	10_0	10_0	10_0
Outcome	1	1	0	1	0
TIR	47%	86%	12%	65%	59%
Count ME Actually visited	16	5	17	6	12
Count RR Elem	3	3	2	3	3
Total RR Elem	3	3	3	3	3
SPP	0.19	0.60	0.12	0.50	0.25
SPR	1.00	1.00	0.67	1.00	1.00
SPF	0.32	0.75	0.20	0.67	0.40
SPF2	0.54	0.88	0.34	0.83	0,63

Table 3 introduces a sample of the aggregated data for the model in Fig. 1. From the total valid observations, 10 observations were set aside for validation purposes. Therefore the data file contains a number of 146 observations, where each observation represents a comprehension questions answered by one subject. The data file is available at: http://granturi.ubbcluj.ro/decision_mining/loguri-en.html. In Table 3, one can see that Subject 1 spent about half of his time evaluating RR elements, fixated all the

RR elements (i.e. recall is 1) but the fixated RR elements were a small sub-set of all the model elements fixated (i.e. precision is 0.19). However, Subject 1 gave the correct answer to the comprehension question. Subject 2 fits better our hypothesis that the correct answer was given because he spent most of his time fixating RR elements, fixated all RR elements, and just a small number of model elements outside the RR. Subject 25 contradicts our hypothesis because he gave the incorrect answer despite fixating all RR elements and spending most of his time looking at RR.

In order to validate H1 and H2 we will use the series for the variables in the example. The sample data summary is introduced in Table 4. We first perform a simple correlation analysis of the dependent variable Outcome with the independent variables SPP, SPR, SPF, SPF2 and TIR. The result, introduced in Table 5, shows that there is some limited correlation between the variables.

Table 4. Observation data summary

Variable	*F*	*F2*	*SPP*	*SPR*	*TIR*
Sample size	146	146	146	146	146
Arithmetic mean	0.5742	0.5742	0.4745	0.6591	0.5729
Standard deviation	0.2214	0.2214	0.2547	0.2687	0.2839

Table 5. Simple correlation between dependent and each independent variables

Variable	*F*	*F2*	*SPP*	*SPR*	*TIR*
ANOVA F-ratio	28,247	29.650	17.290	21.446	17.964
ANOVA Significance	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001
Simple correlation r	0.405	0.413	0.327	0.360	0.333

The ANOVA test in Table 5 shows there is a significant difference in the distribution of the independent variable for correct and incorrect answers. We also performed simple regression analysis (last row of Table 5) to find out if the independent variables are associated with the dependent variable Outcome. The ANOVA analysis, the simple correlation and the simple regression analysis in Table 6 shows there is strong evidence that the null hypothesis can be rejected for all independent variables and that, indeed, a higher number of RR elements fixated and the time spent fixating them is connected with a tendency towards giving a correct answer.

Table 6. Simple regression analysis

Independent Variable	***F_{crit} at 95%***	***F-ratio***	***p-value***
F	4.61	28.24	P < 0.001
F2	4.80	29.65	P < 0.001
SPP	3.01	17.29	P < 0.001
SPR	3.64	21.44	P < 0.001
TIR	3.11	17.96	P < 0.001

To evaluate H3 and H4 we performed a) multivariate regression analysis; b) logistic regression to estimate the probability of the binary choice stored as the Output variable; and c) ROC curve and AUC analysis for modeling the influence of independent variables over the probability of giving the correct answer.

Multivariate regression analysis results show that a regression model using SPP (precision) and SPR (recall) explains the outcome variable (correct answer yes/no). We should also stress that positive coefficients for SPP, SPR and TIR implies a tendency to increase probability of a correct answer with a larger number of RR elements examined or more time taken to examine them. This is in line with our hypothesis. However, a model involving both dimensions (SPP and/or SPR plus TIR) does not explain the outcome. Interestingly, multiple regression analysis shows the F-measure is slightly better than F2 measure in a model also involving TIR.

Table 7. Best multivariate regression for Outcome dependent variable

Independent variables	*Coefficient*	*Std. Error*	$r_{partial}$	*t*	*P*
(Constant)	0.2253				
SPP	0.4232	0.1362	0.2515	3.108	0.0023
SPR	0.4758	0.1290	0.2946	3.687	0.0003

Table 8. Multivariate regression for Outcome dependent variable considering both dimensions under scrutiny

Independent variables	*Coefficient*	*Std. Error*	$r_{partial}$	*t*	*P*
(Constant)	0.3123				
SPP	0.2581	0.2976	0.07234	0.867	0.3872
TIR	0.3105	0.2669	0.09682	1.163	0.2467

The coefficients for the Nagelkerke's R^2 and R^2-adjusted coefficients comparing multiple models are introduced in Table 9 showing that a considerable share of variance is explained by the independent variables. This analysis also shows that variations in the Outcome variable are best explained by the SPP – SPR model.

Table 9. R2 and R^2-adjusted coefficients for multivariate analysis

Independent Variables	R^2	R^2*-adjusted*
SPP, SPR	0.1847	0.1733
SPP, TIR	0.1156	0.1032
SPR, TIR	0.1680	0.1564
F, TIR	0.1649	0.1532

Table 10. Logistic regression equation for dependent variable Outcome and independent variables Precision and Recall

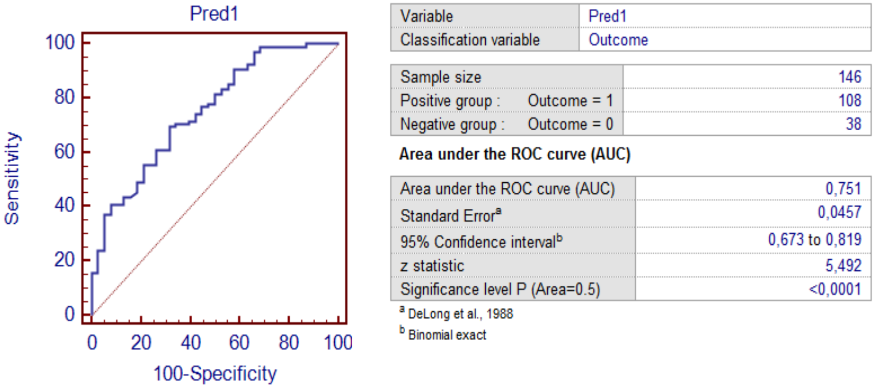
Variable	*Coefficient*	*Std. Error*	*P*
Precision	2.52638	0.95005	0.0078
Recall	2.84899	0.90804	0.0017
Constant	-1.7807		

The logistic regression analysis (in Table 10 and Table 11) shows that giving the correct or incorrect answer to the comprehension question can be predicted by: a) Precision and Recall; and b) Recall and TIR. The other combinations of independent variables are not producing statistically relevant forecasts.

Table 11. Logistic regression equation for dependent variable Outcome and independent variables Recall and Percentage of time spent fixating RR elements

Variable	*Coefficient*	*Std. Error*	*P*
Recall	2.71200	0.94914	0.0043
TIR	1.79751	0.81127	0.0267
Constant	-1.5463		

For predicting the probability of giving the correct answer, we used the logistic regression equation based on the coefficients in Table 11. We try to provide a statement like "for a particular combination of SPR $y_{SPR} \in [0, 1]$ and TIR $y_{TIR} \in [1, 100\%]$, there is a z% probability of giving the correct answer to the comprehension question". Some examples of this statement are depicted in Table 12. The results can be interpreted as the effect of the risk factors that only a part of the model elements are fixated for an insufficient time to give the correct answer to the comprehension question.



Variable	Pred1
Classification variable	Outcome

Sample size		146
Positive group :	Outcome = 1	108
Negative group :	Outcome = 0	38

Area under the ROC curve (AUC)

Area under the ROC curve (AUC)	0,751
Standard Error[a]	0,0457
95% Confidence interval[b]	0,673 to 0,819
z statistic	5,492
Significance level P (Area=0.5)	<0,0001

[a] DeLong et al., 1988
[b] Binomial exact

Fig. 5. ROC curve and AUC analysis for SPR-TIR model

Table 12. The effect on probability of giving the correct answer of Recall-TIR pairs

Recall	*TIR*	*Probability of correct answer*
0.40	40%	56,40%
0.40	60%	64,95%
0.50	50%	67,01%
0.90	50%	85,73%
1	100%	95,09%

For validating our outputs, we followed a machine learning approach using a training set. We considered the SPR-TIR pairs (Table 12) to calculate the probability of a correct answer. We used as reference the 50% threshold (therefore if the prediction was a correct answer with 45% probability we expected an incorrect answer). Then, we manually evaluated the 10 observations set aside from the whole data set, alongside the actual answers given by the subjects. We calculated the prediction precision as the number of true positives over the sum of true positives and false positives. The prediction precision yielded a satisfying 70%. Therefore, we argue that our approach is generalizable and can be used to predict whether experts will provide correct answers to process model based comprehension questions.

5 Related Work

We have approached an actively researched field in the last few years (process model understanding) with an emerging technique for direct observation (eye-tracking). Therefore, we build on previous knowledge from both areas. Main related issues in process model understanding approaches were already presented in the first section of the paper [2], [3], [4], [5], [6], [9], [10], [11], [12], [13]. Most research in eye-tracking is focused related to the areas of psychology, medical research, marketing research and human-computer interaction [19], [20], [23]. The last two areas are linked with information systems because they focus on subjects like advertising, package design, interface design, etc. Most work in information systems is linked to web usability related issues. For the eye-tracking experimental design we used [20], and [23] as references. For eye-tracking data analysis we used [24] as a reference.

Eye-tracking research directly applied to process modeling context is discussed in [15]. Currently, a team of researchers at the University of Innsbruck [16] focus on the process of process modeling and try to combine eye-tracking analysis with previous results on how process models are created. This is obviously approaching a different issue than our research. Somewhat connected to our current goals (but closer to our future goals) is the work-in-progress performed by the same group and introduced in [25]. The approach is to correlate the mental effort with the accuracy of a process model related task, but the results still rely on statistics built on indirect observations.

Eye-tracking was used before to evaluate the difference between novices and experts performing some task. In [23] visualizing scan-path differences between expert and novice aircraft inspectors is used as an example, where the expert scan-path is recommended to be followed by the novices to increase their performance. In our experiments we used only experts, but this approach opens up a new implementation opportunity for our findings.

6 Conclusions

In this paper we introduce a task perspective to the discussion on factors of process model comprehension. We formally defined the notion of a Relevant Region and a Scan-Path, as well as precision and recall metrics based on both. We hypothesize those are connected with the way a subject inspects a process model, and eventually with the performance in giving correct answers. We conducted an experiment using eye-tracking equipment involving 26 expert modelers yielding data on 6 tasks for each of them. Our statistical analysis not only confirms the significance of the hypotheses, but also highlights the predictive power of defined independent variables.

Some limitations of our research are: a) the focus on model structure understanding; and b) that we investigate understanding of a part of the model. To focus on structure we eliminated task labels, and it could be argued that those could play a role in model understanding. But, on one hand, correctly understanding labels is highly subjective. On the other hand, failing to understand the model structure will directly result in poor performance. To address the second limitation, we argue that ‘divide et impera’ strategy can be used in model understanding. Therefore same approach as in our comprehension questions can be applied first to understanding small parts of the model and later to piecing together those parts.

In future research we aim to further investigate the process of model comprehension based on eye-tracking. One direction is to investigate the notion of structuredness in more detail due to its importance for process model comprehension reported in prior research. Also, we will focus on giving a quantitative measure to the difference between novice and expert process modelers when it comes to model comprehension. Our approach can also be used to determine the influence of task (comprehension question) difficulty on the cognitive process of a process model reader. And last, but not least, given a large number of investigated subjects it should be possible to mine some cognitive model exploration patterns (a process of examining a process model) that will ensure a greater probability of understanding the model.

Acknowledgement. This research was supported by Human Resources Development Operational Program through the project Transnational Network of Integrated Postdoctoral Research in the Field of Science Communication, Capacity Building (Post-doctoral School) and Scholarship Program (CommScie) POSDRU/89/1.5/S/63663.

References

1. Recker, J., Rosemann, M., Green, P.F., Indulska, M.: Do Ontological Deficiencies in Modeling Grammars Matter? MIS Quarterly 35(1), 57–79 (2011)
2. Moody, D.L.: The “Physics” of Notations: Toward a Scientific Basis for Constructing Visual Notations in Software Engineering. IEEE Trans. Software Eng. 35(6), 756–779 (2009)
3. Dumas, M., La Rosa, M., Mendling, J., Mäesalu, R., Reijers, H.A., Semenenko, N.: Understanding Business Process Models: The Costs and Benefits of Structuredness. In: Ralyté, J., Franch, X., Brinkkemper, S., Wrycza, S. (eds.) CAiSE 2012. LNCS, vol. 7328, pp. 31–46. Springer, Heidelberg (2012)
4. Reijers, H.A., Mendling, J.: A Study into the Factors That Influence the Understandability of Business Process Models. IEEE Transactions on Systems, Man, and Cybernetics, Part A 41(3), 449–462 (2011)
5. Figl, K., Laue, R.: Cognitive Complexity in Business Process Modeling. In: Mouratidis, H., Rolland, C. (eds.) CAiSE 2011. LNCS, vol. 6741, pp. 452–466. Springer, Heidelberg (2011)

6. Melcher, J., Mendling, J., Reijers, H.A., Seese, D.: On Measuring the Understandability of Process Models. In: Rinderle-Ma, S., Sadiq, S., Leymann, F. (eds.) BPM 2009. LNBIP, vol. 43, pp. 465–476. Springer, Heidelberg (2010)
7. Kiepuszewski, B., ter Hofstede, A.H.M., Bussler, C.: On structured workflow modelling. In: Wangler, B., Bergman, L.D. (eds.) CAiSE 2000. LNCS, vol. 1789, pp. 431–445. Springer, Heidelberg (2000)
8. Laue, R., Mendling, J.: Structuredness and its significance for correctness of process models. Inf. Syst. E-Bus. Manage 8(3), 287–307 (2010)
9. Krogstie, J., Sindre, G., Jørgensen, H.D.: Process models representing knowledge for action: a revised quality framework. EJIS 15(1), 91–102 (2006)
10. Gemino, A., Wand, Y.: A framework for empirical evaluation of conceptual modeling techniques. Requir. Eng. 9(4), 248–260 (2004)
11. Mendling, J., Strembeck, M., Recker, J.: Factors of process model comprehension - Findings from a series of experiments. Dec. Supp. Syst. 53(1), 195–206 (2012)
12. Weidlich, M., Mendling, L., Weske, M.: Efficient Consistency Measurement Based on Behavioral Profiles of Process Models. IEEE Trans. Software Eng. 37(3), 410–429 (2011)
13. Weidlich, M., Polyvyanyy, A., Mendling, J., Weske, M.: Causal Behavioural Profiles – Efficient Computation, Applications, and Evaluation. Fundamenta Informaticae 113(3-4), 399–435 (2011)
14. Decker, G., Mendling, J.: Process instantiation. Data Knowl. Eng. 68(9), 777–792 (2009)
15. Hogrebe, F., Gehrke, N., Nüttgens, M.: Eye Tracking Experiments in Business Process Modeling: Agenda Setting and Proof of Concept. In: Proc. of EMISA, pp. 183–188 (2011)
16. Pinggera, J., Zugal, S., Weidlich, M., Fahland, D., Weber, B., Mendling, J., Reijers, H.A.: Tracing the Process of Process Modeling with Modeling Phase Diagrams. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) BPM Workshops 2011, Part I. LNBIP, vol. 99, pp. 370–382. Springer, Heidelberg (2012)
17. Yoon, D., Narayanan, N.H.: Mental imagery in problem solving: An eye tracking study. In: Eye Tracking Research & Application: Proceedings of the 2004 Symposium on Eye Tracking Research & Applications, vol. 22(24), pp. 77–84 (2004)
18. Salvucci, D.D., Goldberg, J.H.: Identifying fixations and saccades in eye-tracking protocols. In: Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, pp. 71–78. ACM (2000)
19. Granka, L.A., Joachims, T., Gay, G.: Eye-tracking analysis of user behavior in WWW search. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 478–479 (2004)
20. Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., Van de Weijer, J.: Eye tracking: A comprehensive guide to methods and measures. OUP Oxford (2011)
21. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press, Addison-Wesley, New York (1999)
22. Hosmer, D., Lemeshow, S.: Applied logistic regression. Wiley-Interscience Publ. (2000)
23. Duchowski, A.: Eye Tracking Methodology: Theory and Practice. Springer (2007)
24. Salvucci, D.D., Anderson, J.R.: Automated eye-movement protocol analysis. Human-Computer Interaction 16(1), 39–86 (2001)
25. Zugal, S., Pinggera, J., Reijers, H.A., Reichert, M., Weber, B.: Making the Case for Measuring Mental Effort, `http://bpm.q-e.at/wp-content/uploads/2012/09/eessmod_2012.pdf`