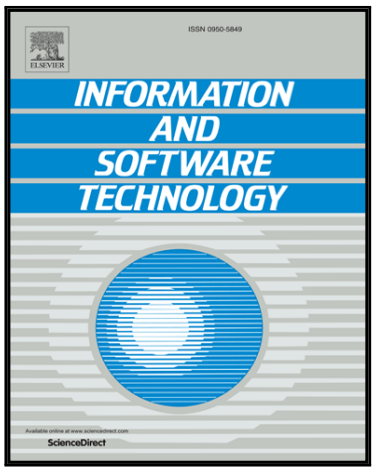


Task-Specific Visual Cues for Improving Process Model Understanding

Razvan Petrusel , Jan Mendling , Hajo A. Reijers

PII: S0950-5849(16)30117-3
DOI: 10.1016/j.infsof.2016.07.003
Reference: INFSOF 5744

To appear in: *Information and Software Technology*

Received date: 11 February 2016
Revised date: 30 June 2016
Accepted date: 11 July 2016

Please cite this article as: Razvan Petrusel , Jan Mendling , Hajo A. Reijers , Task-Specific Visual Cues for Improving Process Model Understanding, *Information and Software Technology* (2016), doi: 10.1016/j.infsof.2016.07.003

Task-Specific Visual Cues for Improving Process Model Understanding

Razvan Petrusel
Faculty of Economics and Business Administration, Babes-Bolyai University, Teodor Mihali str. 58-60, 400591, Cluj-Napoca, Romania Tel. +40 740845115
E-mail: razvan.petrusel@econ.ubbcluj.ro

Jan Mendling
Institute for Information Business, Wirtschaftsuniversität Wien, Welthandelsplatz 1, 1020 Wien, Austria Tel. +43 1313364365
E-mail: jan.mendling@wu.ac.at

Hajo A. Reijers
Department of Computer Science, VU University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, Netherland Tel. +31 402473629
E-mail: h.a.reijers@vu.nl

Abstract:

Context: Business process models support various stakeholders in managing business processes and designing process-aware information systems. In order to make effective use of these models, they have to be readily understandable.

Objective: Prior research has emphasized the potential of visual cues to highlight relevant matters in models such that stakeholders can use them more efficiently. What prior research does not explain is in how far visual cues can be customized to specific understanding tasks and how this influences cognition.

Method: In this paper, we address these questions with an experimental research design, in which we use eye-tracking equipment to capture how process experts use models to answer comprehension questions. As a treatment, we designed two manipulations of the secondary notation, namely coloring and layout, to direct attention to the elements relevant for the specific tasks.

Results: Our results indicate that both manipulations improve both eye-tracking-based measures and performance measures such as duration and efficiency, with color having the stronger effect.

Conclusions: Our findings lay the foundation for novel features of process modeling tools that provide modifications of secondary notation in response to specific user queries. More generally,

our research emphasizes the importance of the relevant region associated with a particular model understanding task.

Keywords:

business process model understanding; visual cues; scoping task specific model elements; process model relevant region; color process model elements.

Classification JEL Codes
L20; M10; L86

1. Introduction

Business process models – or *process models* for short – are important aids for the management of business operations, designing process-aware information systems, and for making decisions during the execution of a process. Process models describe, in essence, the activities that are supposed to be executed as part of the process along with their temporal and logical relationships [1]. They are typically depicted as graph-based diagrams, in which activities, events and gateways are connected with directed arcs. Several process modeling languages exist, with the Business Process Model and Notation (BPMN) [2] being the most widely used one.

It is crucial that process models can be readily understood by the stakeholders for which they are intended. This observation has triggered a stream of research that investigates factors that influence the understandability of a process model, including model complexity [3], notational aspects [4], or the expertise of the model reader [5]. In this context, understanding is typically measured using recall tasks, retention tasks or problem-solving tasks and the corresponding share of correctness of answers, task duration and efficiency. Many of the understanding factors identified by prior research have the weakness that they cannot be easily manipulated. A good mechanism for improving understanding is, however, *secondary notation* [6], also called concrete syntax [7]. Secondary notation comprises all changes to the visual appearance of a model that do not change its logical structure and its semantical interpretation. The benefits of coloring process models as a specific mechanism of secondary notation has been emphasized in [8] and empirically tested in [9]. What prior research, however, does not explain is in how far visual cues can be customized to specific understanding tasks and how this influences cognition.

In this paper, we address these questions with an experimental research design that uses eye-tracking equipment to capture how process experts use models to answer comprehension questions. As a treatment, we designed two manipulations of the secondary notation, namely coloring and layout, to direct attention to the elements relevant for the specific tasks. Our results indicate that both manipulations improve both eye-tracking-based measures and performance

measures, with color having the stronger effect. Our research has implications for process modeling research by clarifying the benefits of task-specific secondary notation and for the general field of conceptual modeling by highlighting the usefulness of the concept of a relevant region. These findings are also interesting from a practical angle since process modeling tools can adapt corresponding features.

The rest of the paper is structured as follows. Section 2 summarizes the research background of process model comprehension and formulates our hypotheses. Section 3 describes our research design and explains how the experiment was operationalized. Section 4 presents the results of our statistical analysis. Section 5 discusses implications of our research for research and practice, as well as limitations. Section 6 concludes the paper with a summary and an outlook on future research.

2. Background

In this section, we describe the background of our research. First, we revisit insights from prior works on process model understanding and visual cognition. Then, we establish the notion of a Relevant Region. Finally, we formulate our research hypotheses.

2.1 Process Modeling and Visual Cues

Process model comprehension is the focus of a considerable number of papers [3], [4], [8], [9], [10], [11], [12], [13], [14]. Typically in these works, an empirical investigation is conducted using controlled experiments of various designs. At the core of those experiments are comprehension questions. Such questions seek to assess if the subject can correctly determine the relationship between two activities in a process model (e.g. concurrency, exclusiveness, sequence, etc.). In general, model comprehension is associated with performance. This is typically measured as *correctness* (i.e. to which degree answers to comprehension questions are correct) and *task duration* (i.e. how fast the answer is given) [14], [15].

Prior research has investigated a number of factors that influence correctness and task duration. It has been found that the more complex process models are, the more they become difficult to understand [10], [13], [16], [17], [18], [19]. In this context, *model complexity* has been measured using various metrics [20], [21] related to model size and model structure [22], [23], [24], [16]. Also, most researchers agree on the benefits of structuredness for model understanding, which means that any split gateway has a matching join [25], [26], [27], [28], [29], [30]. A second rule of thumb associated to process model understanding postulates the benefits of expertise [5], [9], [11], [31]. However, this rule has been nuanced in various papers as different types of expertise were identified and evaluated [5], [31], [32].

From a cognitive perspective, complexity has been described using two dimensions: *stimulus complexity*, which refers to the richness of the content, and *task complexity*, which relates to the mental information processing required to solve the task. Expertise plays a major role in dealing with such complexity in general. Experts also perform better than novices in process model comprehension tasks, which is attributed to a better management of task complexity [3], [9], [12], [13]. Previous research has shown the benefits of the secondary notation on process model understanding [8], [9], but without explicitly linking the concept to stimulus or task complexity.

2.2 Process Modeling and Visual Cognition

The mentioned works from prior research have in common that they refer to cognitive research as a theoretical basis for explaining the empirical connections. The most prominent theory in this context is cognitive load theory (CLT). This theory postulates that the human information processing is limited by the capacity of the working memory [33]. Applications of the CLT are, among others, concerned with the design of instructional methods in such a way that the limited workload memory capacities are utilized in the most effective manner when new knowledge is acquired. Research on diagrams has found chunking to be effective in general [34] and also for process modeling in particular [11]. In order to support this, instructional methods should

decrease the mental effort [35] in such a way that readers can easily chunk the process model, such that the relevant information can be directly spotted.

The predominant theoretical paradigm in visual languages research, Cognitive Dimensions of Notations, is limited when it comes to evaluating the design of visual notations [36]. One of the reasons is that dimensions lack clear operationalization of evaluation procedures and metrics. The effectiveness of diagrams can be influenced on the semantic, visual representation and sentence levels [36]. Moody also recommends several principles for designing effective visual notations [36]. The principles for reducing diagrammatic complexity and visual expressiveness can be instantiated both at visual representation level as well as sentence level by techniques that direct attention to a subset of model elements. In this way, excessive graphic complexity can be addressed by increasing human discrimination ability without revising notation semantics and syntax.

For guiding the attention of a model reader effectively and thus reducing cognitive load, research has investigated many techniques, a class of which are *visual cues*. Of particular interest in relation to process models are two cueing techniques. First, *coloring* has been used to highlight the matching splits and joins in a Petri-net [9]. Though improving task duration and correctness for novices, this kind of cue showed no improvement for experts. Similar effects have been reported also for UML diagrams [37]. On the other hand, Moody [36] and the UML specification, while acknowledging the sensitivity of coloring, recommend against its use. We argue that such types of coloring that reduce the perceived complexity in relation to a specific tasks are likely to improve comprehension performance. Therefore, we formulate:

Proposition 1: Task-specific coloring improves comprehension performance.

Second, the *graphical layout* of a model has been studied as a factor of comprehension [38]. Different layout metrics called *aesthetics* have been defined in order to help creating an optimal layout. For process models, the notion of structuredness integrates both a logical and an aesthetic perspective [25], [26]. Sharif [39] suggests though that graph aesthetics should be sometimes

violated to achieve a better comprehension. This argument too points to the potential of providing a layout that is specifically aligned with a specific task, for instance by repositioning elements to fit the tasks, such that the perceived complexity is reduced. We therefore formulate:

Proposition 2: Task-specific layout improves comprehension performance.

At this point, a note of caution has to be made, since works exist that question any potential benefits. For instance, [40] find that task duration is reduced if highlighting guidance was 80% to 100% accurate. However, a drop of accuracy below a threshold of 60% reduced the practical usefulness. Similar results were reported for map reading [41]. Even though not all performance dimensions might be improved by visual cues, there is a broad consensus on their benefits.

2.3 Scoping Visual Cues for Process Model Comprehension Tasks

In order to design visual cues for a process model comprehension task effectively, we need to investigate which part of the model is relevant for that task. Traditional complexity metrics might be misleading here, since a *simple* question how two neighboring activities are related can be asked with reference to a *complex* model. Prior research defined the notion of a Relevant Region (RR) [42], and revealed an empirical connection between the number of Relevant Region elements fixated by the reader and correctness of answering comprehension questions.

The concept of RR uses the definition of a process model, the notion of a path, and the notion of a dominator from [43]. A generic process model is a tuple PM = (N, F, l) where N is the set of nodes, which is further partitioned as $N = S \cup E \cup A \cup G$, where S is set of the start events; E the set of end events, A the set of activities and G the set of gateways. F is the set of arcs, defined as $F \subseteq N \times N$. The function $l: G \rightarrow \{AND, OR, XOR\}$ maps gateways to corresponding label types AND, OR, and XOR.

A path from a node n_1 to a node n_k, $n_1 \rightsquigarrow n_k$, is a non-empty sequence of nodes such that $(n_1, n_2),\ldots,(n_{k-1}, n_k) \in F$. For two nodes x and y, we define x = dom(y) as a dominator of y if and only if for all paths from a unique start event $s \rightsquigarrow y$ it holds that $x \in (s \rightsquigarrow y)$.

In case there are several start events, the notion of a dominance frontier [44] extends the notion of dominator as the set of all nodes y such that x dominates a predecessor of y but does not strictly dominate y. Inspired by the notion of a start join in [45], we define the relevance frontier RF(Y) of a set of nodes y_i, $\in Y \subseteq N$ that are dominator to all $x_i \in X \subseteq N$, and which have no other dominators on their paths to y_i in case such a path exists.

Based on these notions, a Relevant Region concerning tasks y_i, $\in Y$ is denoted as, $RR(Y, PM) \subseteq N$ and defined as:

$$RR(Y, PM) = \{n \in N \mid n \in x_i \rightsquigarrow y_i \text{ with } x_i \in RF(Y) \text{ and } y_i \in Y\}.$$

To give an intuition of RR, let us consider Figure 1A, as well as the comprehension question "If G is executed, must J be executed in the same case?". The comprehension question refers to two tasks, therefore the first elements of the RR are tasks labeled 5 and 6. The path from *start* node to node 5 runs via the gateways labeled 1 and 2. The path from *start* to node 6 contains the gateways labeled 1 and 3. Therefore, the dominating node is 1. There is only one path from node 1 to node 5, which includes gateway 2. Similarly, gateway 3 is on the only path from node 1 to node 6. Therefore, RR = {1, 2, 3, 5, 6}.

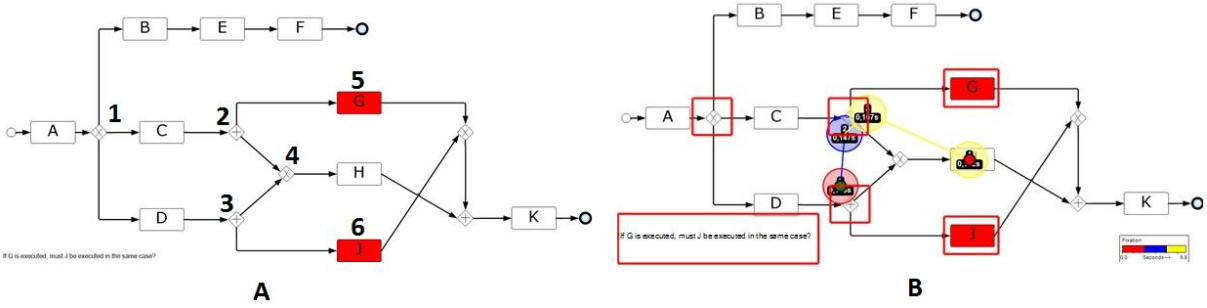


Figure 1 – Example of (A) benchmark model, and (B) model annotated with visual eye-tracking output with Fixations and the Area of Interest for Relevant Region elements

By looking at the comprehension question, one will reason that in order to get node 5 executed, the first XOR-split (node 1) has to activate middle branch towards node 2. In that case, the lower branch will not be executed. Node 2 (AND-split) will pass control to node 5 and to node 4 (XOR-join). Once at node 4, there is no way back to node 6. The same reasoning applies if node 1 activates the lower branch towards node 3. Therefore, nodes 5 and 6 will never be executed in the

same instance. So, the answer to the comprehension question is 'No'. The RR notion supports the idea that only gateways 1, 2 and 3 have to be investigated in order to arrive at the answer to this comprehension question, all other model elements being irrelevant.

This approach is tailored for questions on behavioral relations between model activities and assumes that the model is sound. Soundness implies no deadlocks, no livelocks and proper synchronization for any concurrent branches [46]. Therefore, investigating only the path from the start node to the input tasks will provide all the necessary information for assessing the relationship between the tasks.

2.4 Measures of Comprehension Performance

Comprehension effectiveness (or performance) in relation to conceptual models and visual notations can be measured in a variety of ways [36], [15]. The classical measures are *correctness* of answers to a particular set of comprehension questions (also referred to as accuracy), *task duration* for answering questions (known as completion time), and *efficiency* as correctness divided by task duration. Even though these measures are widely used, they hardly provide insights into how persons approach a comprehension task.

Cognitive effectiveness of visual representations was empirically evaluated using indirect observation methods, mainly surveys [3], [47]. Since we aim to investigate visual cues, it is of utmost importance to study visual cognition more directly. Eye-tracking technology provides effective devices to support research based on direct observation, which closely reflects the cognitive processes of a subject [48], [49]. Medical and psychological research has shown that there is a so-called eye-mind relationship, which means that accurate perception requires persons to fixate the object with the eyes and to focus their minds on it (i.e. this is commonly named *attention*) [50], [51]. Therefore, the easiest way to detect the object of attention is to measure on which point someone's eyes are fixated. Eye-tracking is a technique that pinpoints the object of interest of a subject based on the eye mechanics. Therefore, it is superior to think-aloud protocols that may influence a person's thinking process and that only provide qualitative data [52]. Despite these advantages, eye-tracking studies in process modeling are scarce with [11], [42] and [53]

being the few existing studies. A survey on eye-tracking studies in software engineering is reported in [54].

Eye-tracking is well suited to study the significance of task-specific visual cues. The eye-tracking observation method records several metrics that are interesting in this context [48], [49]: the *number* and the *duration* of each fixation, which is the pause of eye movements on a specific area of the visual field. In Figure 1B, each of the squares around the RR gateways defines a so-called Area of Interest (AoI). AoIs are defined during data analysis in order to link a fixation on some screen coordinates to a certain model element. One can see that the subject had four fixations (shown as circles with numbers in the middle) on three elements of the model, out of which two were RR elements (node 2 twice and node 3 once).

2.5 Research Hypotheses

Above, we have discussed the potential benefits of task-specific visual cues for comprehension performance. Our mechanism for *tying* the visual cues to the process model is based on the Relevant Region, such that only relevant elements are highlighted. Our mechanisms for *highlighting* are color and layout modifications on the level of the secondary notation of the process model. We operationalize performance based on correctness, task duration, efficiency, and number and duration of fixations.

For the Color manipulation of the secondary notation, we formulate the following hypotheses:

H1.1 – Manipulating secondary notation by **Coloring** relevant model elements, **decreases** the number of **Fixations** on the model.

H1.2 – Manipulating secondary notation by **Coloring** relevant model elements, **decreases** the **Total Duration of Fixations** on the model.

H1.3: Manipulating secondary notation by **Coloring** relevant model elements **improves Correctness** of comprehension question answers**.**

H1.4: Manipulating secondary notation by **Coloring** relevant model elements **decreases** the **Duration** of comprehension tasks**.**

H1.5: Manipulating secondary notation by **Coloring** relevant model elements **improves Efficiency** of comprehension tasks**.**

For the Layout manipulation of the secondary notation, we formulate the following hypotheses:

H2.1 – Manipulating secondary notation by **Layout decreases** the number of **Fixations** on the model**.**

H2.2 – Manipulating secondary notation by **Layout decreases** the **Total Duration of Fixations** on the model**.**

H2.3: Manipulating secondary notation by **Layout improves Correctness of comprehension question answers.**

H2.4: Manipulating secondary notation by **Layout decreases** the **Duration** of comprehension tasks**.**

H2.5: Manipulating secondary notation by **Layout improves Efficiency** of comprehension tasks.

In the next section, we will consider the design and execution of an experiment to investigate these hypotheses. This includes in particular the exact description of *how* color and layout are manipulated.

3. Research design and execution

In designing, implementing, analyzing and reporting the results of our experiment, we used the recommendations in [55], [56]. Our experimental design uses the two factors, color and layout modification, in comparison to a control group of process models without color and in standard layout. We investigated the influence of these factors on model understanding using a within-subject design. This section describes the factors, objects, subjects, experiment variables and the data collection procedure.

3.1 Experimental factors

The experiment factors in our experiment have two levels. First, we color highlight those gateways that are relevant for the comprehension question being asked. The benchmark models

use the standard notation without color. For the treatment models, RR model elements are colored in red. Second, we change the layout of treatment models separately. The control models use the standard notation with gateways of same size and positioned from left to right. For the treatment models, the RR model elements are bigger in size compared to the rest of the gateways and repositioned close to each other while keeping the overall model structured.

3.2 Experimental objects

The *experiment objects* are 16 models created using the Business Process Model and Notation (BPMN). We started out with 8 different benchmark models; for each of these we created a second variant implementing a treatment to be tested. Therefore, coloring and layout were tested on 4 different models. Models were introduced in the same sequence for all participants. We used a balanced mix of models from academia and industry.

Model understanding is tested by asking participants comprehension questions, inspired by the ones in [3], [42]. One question was asked for each model. The comprehension questions test understanding of *sequence* (e.g. "Can M be executed if J was not executed before?"), understanding of *concurrency* (e.g. "If H is executed, will K always be executed in the same case?"), and understanding of *exclusiveness* (e.g. "If G is executed, must J be executed in the same case?"). The correctness of these relationships can be formally checked using behavioral profiles [57].

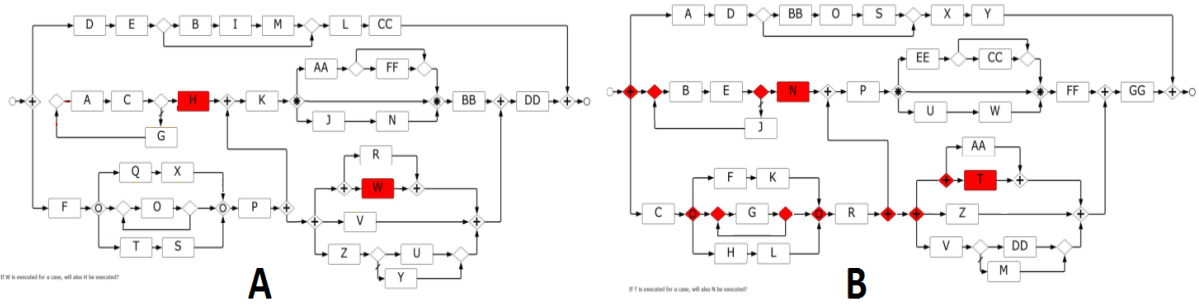


Figure 2 – Model 1, with variants: A) benchmark model, B) Relevant Region highlighted with color.

Figure 2 shows an example of the color treatment. One can see that the size and the layout of the models are identical. In this instance, the only difference is that 10 of the gateways belonging to the respective Relevant Region in model B are colored red.

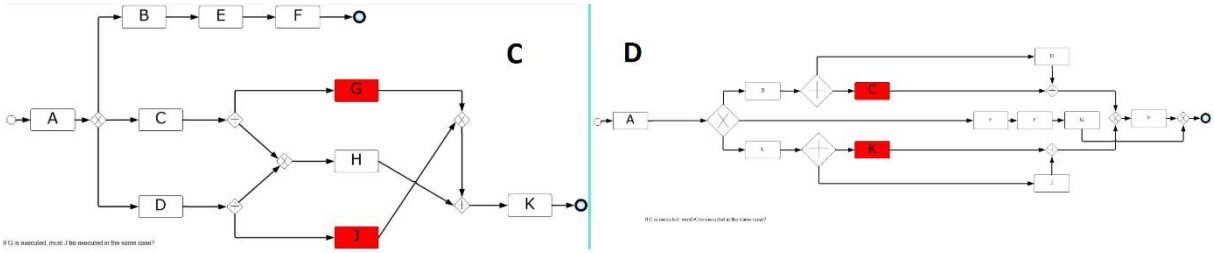


Figure 3 – Model 7, with variants C) base model, and D) layout treatment

Figure 3 shows an example of the Layout treatment. While it is essentially the same model, the RR elements are repositioned in the center of the model. Resizing makes the three RR gateways much bigger than the other gateways.

3.3 Experimental subjects

The *experiment participants* were 75 experienced modelers from industry and academia. We aimed for a medium-to-high level of BPMN knowledge by using a combination of experts (both in BPMN and other process modeling notations) and several subjects with an intermediate level of expertise (i.e. Ph.D. and master students). We also aimed for a balanced mixture of professionals and academics. In effect, 44% of subjects have an industrial position with companies such as Camunda Services GmbH Berlin Germany, Perceptive Software Apeldoorn The Netherlands, and Signavio GmbH Berlin Germany; 56% of our subjects are in academia, notably at universities such as HPI Potsdam, HU Berlin, TU Eindhoven, UBB Cluj-Napoca, WU Vienna. Experiments were conducted with subjects on-site in Austria (Vienna), Germany (Berlin, Potsdam), The Netherlands (Apeldoorn, Eindhoven), Romania (Cluj-Napoca), and Switzerland (St. Gallen). All subjects participated voluntarily; no reward was provided.

To assess the expertise of the participants, several questions were asked. As can be seen in Table 1, the participants self-reported their familiarity with BPMN on average as rather high (both in terms of general knowledge and reading BPMN models). They have read on average about 40 models in the last year. From the 75 subjects, 3 data points were eliminated due to memory effects or problems with eye-tracking recording. Accordingly, 72 subjects were used in the analysis.

Table 1 – Expertise-related variables of the participants

Variable name	*Sample*	*Min*	*Max*	*Mean*	*Std. Dev.*
Familiarity with BPMN	72	1	5	3.43	1.173
Familiarity with reading BPMN models	72	1	5	3.57	1.046
Number of any kind of process models read in the last year	72	0	200	38.50	41.702

3.4 Experimental procedure

As for the ***experimental procedure***, a group of 4 slides were used to operationalize each comprehension question: Slide 1 shows the comprehension question; Slide 2 shows the model; Slide 3 asks the subject to provide a true-or-false answer by clicking on radio buttons; while Slide 4 asks the subject to provide a self-evaluation of the confidence in the provided answer (radio buttons ranging from 1 to 5). Each participant walked through a demo question based on a sample model. Additional explanations were given to the subject (e.g. that there was no time limit, that they were allowed to reasonably move their heads, etc.).

The experimental data collection was conducted in four phases

1) Calibration followed by a set of 8 questions (all benchmark models). After Part 1 there was a 1 minute break.
2) A re-calibration followed by a second set of 8 questions (questions 5-8 were the Color treatment models and questions 9-12 were the Layout treatment models).
3) Still using the software tool, participants answer additional questions that tested the memory effect (i.e. “Did you recognize the same model being shown twice?”); self- assessment expertise questions; and theoretical questions related to the topics under investigation (e.g. “If two activities are concurrent, then they must be executed at the same time?”).
4) At the end of the experiment, we gathered qualitative information from each participant by an informal discussion by asking questions like: “which visual cueing technique you prefer?”, “why is that?”, “did you trust the highlighted elements as relevant?”, “any personal residual impressions?”, etc.

3.5 Experimental instrumentation

The main ***experimental instrument*** was the S2 eye-tracking system provided by Mirametrix (www.mirametrix.com). The Mirametrix software suite consists of three tools. First, the EyeMetrix Design offers a slide-show experiment implementation. Second, the EyeMetrix Record is concerned with runtime data collection. Third, the EyeMetrix Analyze facilitates raw data aggregation and eye-tracking data export.

The hardware component is a binocular, video-based remote eye-tracker. This means cameras are used to capture eye movements. The tracking method is 'bright pupil', which means the eyes are illuminated by infrared light thus making them brighter than the rest of the eye. Then the cameras film those bright spots and the software converts the video into screen coordinates. The cameras are built into an independent device that is placed under an LCD screen. The screen resolution was set to 1280*1024 on a 19 inch screen. The eye-tracker captures data at a rate of 60 Hz, with an average accuracy between 0.5 and 1 degree, which translates to an error between 15 and 30 pixels. The eye-tracker compensates for head movement during the study, i.e. a subject's head is free or his eyes do not have to be focused on the screen all the time. The calibration was done using 9 points.

3.6 Experimental data

Raw ***experimental data*** collected using EyeMetrix Record is available both in video format with screen capture including audio and in numerical format. The latter include gaze position, fixation sequence data, and answers to experimental questions. Fixations on selected model elements are discriminated from all the fixations by creating Areas of Interest (AOI). AOIs were kept consistent at same size and positioning around elements across benchmark and treatment models. The fixation duration threshold set in the analysis phase was the eye-tracker default settings of 0.1 seconds (same threshold was used in [41], [58], [59], [60]), with a distance threshold of 40 pixels. The fixation duration threshold is the briefest possible pause of eye-movements while the distance

threshold is the furthest distance between screen coordinates. Both are used in order to discriminate between a saccade and a fixation.

The *experimental variables* include model complexity, expertise, mental effort and response variables related to performance. To operationalize *Model Complexity*, we counted the number of control-flow elements in the model (No of Control-Flow Elements). The Relevant Region Size (RRS) measures comprehension task complexity as the number of relevant gateways in the model. For assessing *Expertise* we computed a single variable that is the average of the participant's self-evaluation on: the Familiarity with BPMN in general; the Familiarity with reading BPMN models; and the Number of any type of process models read in the last year. To normalize the latter over an interval of 1 to 5, user responses were stored as whole numbers and then divided by 40, the rough mean of the sample data.

We need to measure the *Mental Effort* involved in process model understanding. In line with the findings reported in [61], we rely both on fixations count and fixation duration metrics. In our particular setting, the variable Total Count of Fixations on the model (Total Fixations, for short) is a count of eye pauses on various elements and it directly measures the cognitive effort of the subject. Variable Total Duration of Fixations measures the total time the eyes of the reader are fixated by adding the duration of each individual fixation.

We used three *Response Variables*. The most commonly used metric in process model understanding research [3], [9], [12], [21] is the correctness of answering the comprehension questions (Correctness, for short). Of the 16 questions asked, half had *yes* answers and half had *no* answers. Furthermore, we record Task Duration as the time a participant spends examining a model: it does not include the time necessary for reading or answering the comprehension question. This precise measurement was achieved by placing the question, the model and the answer on three different presentation slides. A third metric is Efficiency calculated as Correctness over Task duration. It captures the ability of the subjects to give the correct answer in a timely manner.

4. Results

In this section, we present the results of our experiment. First, we provide an overview of the data by summarizing descriptive statistics and we screen the data for correlations. Second, we test the introduced hypotheses regarding the impact of the factor levels by pair-wise tests that evaluate the differences between benchmark and the treatment models. As to the hypotheses regarding mental effort, we test for the impact of the treatments on variables Total Fixations and Total Duration of Fixations. As to the hypotheses that focus on understanding performance, we test for the impact of the treatments on the Correctness, Task Duration and Efficiency.

4.1 Descriptive statistics and correlation analysis

The data recorded using EyeMetrix Record was aggregated and summarized using EyeMetrix Analyze and then imported into SPSS for statistical analysis.

As a pre-processing step, observations were filtered such that only the ones that showed less than 10% missing data were kept. Missing data are invalid coordinates that may show up because the subject looked outside the screen or the eye-tracker lost track of the subject's eyes. For example, if the subject moves the head swiftly, there is a slight delay until the pupils are focused again. Furthermore, we eliminated one data point which was an outlier in terms of duration, because the participant took four times as long as the second slowest. Altogether, there were 937 valid observations, with one observation representing one comprehension question answered by one subject based on one model. Table 2 summarizes the mean, standard deviation, and range of all variables.

Table 2 – Descriptive Statistics of the Variables

Variable	N	Mean	Std. Dev.	Range (min to max)	Normality
No of Control-Flow Elements	937	17.95	9.45	6 to 33	NO, p<0,001
Relevant Region Size	937	7.59	2.93	4 to 12	NO, p<0,001
Familiarity	937	2.60	0.93	0.66 to 4.58	NO, p<0,001
Total Fixations	937	52.91	44.05	0 to 271	NO, p<0,001

Total Duration of Fixations	937	10,974	10,008	0 to 63,28	NO, p<0,001
Correctness	937	0.87	0.34	0 or 1	NO, p<0,001
Duration	937	24.54	17.84	2.8 to 111	NO, p<0,001
Efficiency	937	0.06	0.058	0 to 0.36	NO, p<0,001

We need to point out the high percentage of correct answers (87%) as shown in this table, as this impacts our further analysis and findings.

Table 3 – Variable correlation matrix

	No Of Control-Flow elem.	Relevant Region Size	Familiarity	Total Fixations	Total Duration of Fixations	Correctness	Duration
Relevant Region Size	0,769 p<0,001						
Familiarity	-0,005 p=0,877	-0,013 p=0,702					
Total Fixations	0,300 p<0,001	0,327 p<0,001	-0,277 p<0,001				
Total Duration of Fixations	0,278 p<0,001	0,312 p<0,001	-0,301 p<0,001	0,979 p<0,001			
Correctness	0,027 p=0,414	-0,080 p=0,014	0,123 p=0,002	-0,061 p=0,064	-0,066 p=0,042		
Duration	0,330 p<0,001	0,360 p<0,001	-0,217 p<0,001	0,932 p<0,001	0,892 p<0,001	-0,065 p=0,046	
Efficiency	-0,300 p<0,001	-0,356 p<0,001	0,238 p<0,001	-0,554 p<0,001	-0,529 p<0,001	0,401 p<0,001	-0,590 p<0,001

Table 3 shows the correlation matrix. Before we turn to the actual testing of the hypotheses, we screened the data for potential interactions within the variable set. At a first glance, independent variables seem correlated with the dependent variables Duration, Efficiency and Total Fixations. As the participants did well in answering the tasks correctly, we observe insufficient variation to establish proper correlation between Correctness and the other variables. Obviously, Total Fixations is strongly correlated with Duration given that the longer the model is investigated, the more fixations occur. A significant but moderate correlation exists between Total Fixations and the independent variables Familiarity, No of Control-flow Elements and Relevant Region Size.

All these connections are in line with previous research on model understanding on the one hand and with our expectations on the other hand. Therefore, we proceeded with the data analysis.

4.2 Pair-wise test

In order to test the effect of the secondary notation manipulation, we turn to pair-wise t-tests.

First, we checked whether the requirements of the t-test were met. The Kolmogorov-Smirnov test indicates that the assumption of normality for all of the four dependent variables is not met. Therefore, we used the non-parametric Wilcoxon test, which is based on ranks. In essence, this test works with the differences between the observation without and with treatment. For example, this would be the difference between Correctness of a question asked on a model *without* secondary notation manipulation and a model *with* treatment (e.g. by coloring RR gateways). A negative rank would then indicate the model with color having a result that is worse than the one without color. The Z-score then captures to what extent the amount of the negative or positive ranks relates to the overall number of observations. The effect size r can be calculated based on the Z-score divided by the square root of the number of observations N [56], where r = 0.3 indicates a moderate effect.

Table **4** shows that the Color treatments significantly decreases Duration, where the reduction is in the range of 10%. Correctness does not seem affected, but through the reduction of Duration, Efficiency still improves significantly. Mental effort is also significantly lower for the models with a Color treatment, both measured in terms of the number of Fixations (p=0.0013) and Fixation Duration (p=0.0008). The strongest of the noted effects relates to Efficiency with r = 0.37, a moderate effect.

Table **5** shows that the Layout treatments also improves Efficiency, and significantly so (p=0.0002). The decrease of Duration is not statistically relevant when assuming a 95% confidence interval (p=0.0712). Just as in the case for Coloring, mental effort is also significantly reduced, both in terms of significantly less Fixations (p=0.0085) and a lower Fixation Duration (p=0.0019). Of all the significant relations, the effect on Efficiency is the highest with an effect size of r = 0.264, which is a moderate effect.

Table 4 – Wilcoxon pair-wise tests results for Color treatment

Dependent Variable	Treatment Group	Variable Median	Ranks	N	Z	p and r
Correctness	Benchmark Treatment	1 1	Negative Positive Ties	13 23 172	-1.667	**p = 0.096** **r = -0.11**
	Paired differences	0	Total	212		
Duration	Benchmark Treatment	28.47 24.30	Negative Positive Ties	126 86 0	3.039	**p = 0.002** **r = 0.21**
	Paired differences	-2.69	Total	212		
Efficiency	Benchmark Treatment	0.0317 0.0397	Negative Positive Ties	71 134 7	-5.396	**p < 0.0001** **r = -0.37**
	Paired differences	0.01	Total	212		
Total Fixations	Benchmark Treatment	65 50	Negative Positive Ties	123 88 1	3.218	**p = 0.0013** **r = 0.221**
	Paired differences	6	Total	212		
Total Duration of Fixations	Benchmark Treatment	13.14 10.2	Negative Positive Ties	128 84 0	3.340	**p = 0.0008** **r = 0.229**
	Paired differences	1.94	Total	212		

Table 5 – Wilcoxon pair-wise tests results for Layout treatment

Dependent Variable	Treatment Group	Variable Median	Ranks	N	Z	p and r
Correctness	Benchmark Treatment	1 1	Negative Positive Ties	20 20 163	0	**p = 1** **r = 0**
	Paired differences	0	Total	203		
Duration	Benchmark Treatment	17.28 13.84	Negative Positive Ties	118 85 0	1.804	**p = 0.0712** **r = 0.127**
	Paired differences	- 1.99	Total	203		
Efficiency	Benchmark Treatment	0.050 0.069	Negative Positive Ties	76 118 9	-3.755	**p = 0.0002** **r = -0.264**
	Paired differences	0.015	Total	203		
Total Fixations	Benchmark Treatment	37 27	Negative Positive Ties	119 83 1	2.631	**p = 0.0085** **r = 0.185**
	Paired differences	- 3	Total	203		
Total Duration of Fixations	Benchmark Treatment	7.2 5	Negative Positive Ties	120 83 0	3.111	**p = 0.0019** **r = 0.218**
	Paired differences	- 0.91	Total	203		

5. Discussion, implications and threats to validity

In this section, we first discuss the results in the light of the research hypotheses. Then, we highlight implications for research and practice. Finally, we clarify limitations and threats to validity.

5.1. Discussion of Results

We set up this study to evaluate two main hypotheses. First, we hypothesized that visual cues will improve the performance of process model understanding. In this context, we operationalized task-specific visual cues by coloring the relevant model elements (H1) and by modification of the model layout (H2) for the given comprehension question. Performance was measured as: correctness (H1.3/H2.3) and duration (H1.4/H2.4) of answering comprehension questions, as well as efficiency (H1.5/H2.5) that considers both. Second, we hypothesized that the secondary model manipulations will reduce the mental effort of the readers by decreasing their cognitive load. Measured variables linked to mental effort were the Total Fixations (H1.1/H2.1) and the Total Duration of Fixations (H1.2/H2.2).

Table 6 – Overview of Hypotheses and Results

Hypothesis	**Performance Dimension**	**Support**
H1.1 Coloring -> Mental effort	Total Fixations	YES
H1.2 Coloring -> Mental effort	Total Duration of Fixations	YES
H1.3 Coloring -> Performance	Correctness	NO
H1.4 Coloring -> Performance	Duration	YES
H1.5 Coloring -> Performance	Efficiency	YES
H2.1 Layout -> Mental effort	Total Fixations	YES
H2.2 Layout -> Mental effort	Total Duration of Fixations	YES
H2.3 Layout -> Performance	Correctness	NO
H2.4 Layout -> Performance	Duration	NO
H2.5 Layout -> Performance	Efficiency	YES

From Table **6**, it can be seen that all but one of the hypotheses related to coloring find support. The exception is the improvement of Correctness (H1.3). For the hypotheses that relate to Layout, both hypotheses that relate to the mental effort find support (H2.1, H2.2), but performance only

improved in terms of Efficiency (H2.5). This indicates that a lower mental effort is indeed related to both types of treatments, but that performance improvement is most strongly connected to the proposed coloring approach.

5.2 Implications of Reported Findings

The findings reported in this paper have several implications for research and practice.

The presented research informs us on the way how visual cues help modelers and model readers. Task-specific visual cues such as coloring and modification of the layout can be used to improve the efficiency for solving an understanding task. As the models we used in our experiment where bound to the constraints of the computer screen, these effects might even be stronger for big models of more than 100 elements, which are not unusual in practice. It is also worth mentioning that the results of the statistical analysis are in line with the qualitative feedback that we collected after the experiment. By corroborating our findings with the ones reported in [40], it appears that users can make effective use of highlighted regions. In line with [62], users tend to be sensitive if highlighting is perceived to be intrusive. More specifically, modification of layout was reported as less intrusive if the absolute number, or the percentage of RR elements out of the total model elements that are highlighted, is low. If the number of RR elements is large, resizing was reported as distracting, thus leaving color as the only way to go. As much as the high support for performance-related hypotheses in case of the coloring, this observation also supports coloring as the superior technique for providing visual cues.

Our research also emphasizes the merits of eye-tracking for investigating decision making in conceptual modeling. Most prior studies on process model understanding relied on performance data alone. There is a huge potential of using this technology more intensively in future research, for instance to investigate how modelers direct their attention through the process of process modeling [63] and which highlighting techniques appear to be most helpful to increase the quality of the modeling act itself.

The visual design rationale has been criticized to be neglected for various notations in software engineering [36]. Eye-tracking technology is therefore increasingly used in recent years in this area [54]. Our approach to testing cognitive effectiveness of visual improvements to the secondary notation by eye-tracking controlled experiments can be adopted in this area and with different notations. This might inspire also research such as [64] that bridges between software engineering and process modeling.

The approach used in this paper can further inform research on process model repositories. Process querying is a theoretically properly established method, grounded in process model behavior, which calculates process models similarity with an input model fragment [65], [66], [67]. Visualization in model repositories is similar to querying but aimed at graphical depictions of differences between models [68]. Also connected to repositories, methods for automatic changes in structure [69] or in layout of the models [70] were formally defined. As the basic idea in all these areas is similar to our approach, provide an empirical angle for investigating the impact on users.

Our research has also implications for practice. Currently, we are experimenting with facilities that allow a modeler to select two activities such that the tool then dynamically calculates the relevant region. This tool implements algorithms that highlight the relevant gateways and, at this stage, leaves further reasoning to the reader. Commercial tool vendors can easily integrate highlighting techniques for relevant regions based on the definitions provided in this paper, and extend it with further refinements such as recommendations for specific types of problems. This can also be helpful to support system analysts in their interaction with domain experts or teachers that aim to clarify specific behavior in a model to their students. While our experiment was conducted with BPMN models, we believe that the results can be equally applied for tools that use graph-based process modeling languages such as EPCs, Petri Nets or UML Activity Diagrams.

5.3 Limitations and Threats to Validity

There are limitations and threats to the validity of our study that need to be mentioned. Internal validity might be threatened by the imprecision of the eye-tracking system. To mitigate this type of risk, we used large screens as well as models with sufficient white spaces around each element. Also, during analysis we defined an area of interest larger than the element itself, so that we could compensate for the eye-tracking system's constructive imprecision (i.e. a maximum of 0.5 centimeters).

Other internal validity concerns were identified and addressed. To mitigate the threat of inconsistent measurements, we used models with the same size of elements (of course, for the Layout treatment, the Relevant Region elements had to be bigger). We carefully considered the threat of learning effects. First, we started the experiment with a training task such that the subjects could get familiar with the experimental material. Second, we modified the layout between model variants using mirroring as exemplified by [71], relabeled activities, and displayed eight other models before showing the variant. It is well-known from psychology that working memory information, which is needed to solve the tasks in our experiment, is forgotten if not further integrated after 30 seconds [72]. Third, we asked the subjects after the experiment if they recognized the same model being shown twice and if they gave the same answer, remembering the previous answer. Out of the 75 subjects, only one indicated memory effect (therefore, all observations connected with them were removed from the data set). This case highlights that learning effects cannot be fully ruled out, but that it is unlikely that they had a distorting effect on the data. Note that we deliberately selected subjects familiar with BPMN who felt confident and knowledgeable to answer the types of questions we asked. Also threats in terms of fatigue have to be considered. We deliberately designed the experiment to take not too much time. The average completion time was 25 minutes (including answers to the expertise questions). We also had a small break in the middle of the experiment for calibration, which the subjects could also use as a mental refreshment.

To ensure construct validity, we examined the data and found no unexpected correlations. A possible threat to conclusion validity arises from the lack of a direct comparison between the Color and the Layout treatments. Indeed, we used two different sets of models for the two factor levels. However, a direct comparison between the Color and the Layout treatments was out of scope. We only aimed and evaluated if the two treatments had an impact on understanding and mental effort.

To address external validity concerns, we aimed both for a mix of academia and industry participants and geographical spread, in an effort to ensure a high degree of generality to our findings. Also, we used models created with the BPMN given that it is the most widely adopted type of modeling notation in teaching and in industry. To address risks related to the nature of the models, we used models used by other researchers [3] and models from industrial sources in a balanced way. Another aspect is the high correct response rate attained by the expert participants involved in our study, which rendered inconclusive results for the Correctness metric.

The definition of the Relevant Region allows any number of input tasks to be defined. However, as adding an extra task to the input set enlarges the Relevant Region, we experimented only with the sub-set of questions based on two input tasks. As in industry there might be cases when more input tasks are defined, our findings remain valid as the larger problem can ultimately be decomposed into smaller questions involving pairs of the input tasks. On the research side, comprehension questions with two input tasks were used in all up-to-date experiments on process model understanding.

6. Conclusion

The central question in this paper is how business professionals can be better supported when solving problems that require reading process models. We set out to investigate an innovation in the support provided by current modeling tools in the form of visually signaling relevant parts in the model. Specifically, we examined two task-specific visual cueing techniques, i.e. coloring and modification of layout of model elements. Our experimental evaluation shows a clear impact of

these two techniques on the ability of the study's participants to provide correct answers in a timely manner, as well as on the mental effort required to read process models. The findings presented in this paper are also in line with previous research in other domains (e.g. modeling, geography, science concepts teaching, etc.), which postulate a better understanding if *visual* (preferably *dynamic*) cueing is employed. Therefore, we argue that the extension of process modeling tools with support for a dynamic interaction with its users is both desirable and effective. We started to take the first steps towards such an implementation ourselves, by contributing to the open-source software WoPeD.

The findings of this paper may spark future work considering the benefits brought by directing the reader's perception on the model. Clearly, the secondary notation manipulations need not to be confined to the Relevant Region as defined on this paper. One may also wish to pursue other progressive or dynamic forms of attention guidance that may further increase model comprehension. A concrete suggestion that was made by various participants in our evaluation, for example, was to gray out irrelevant parts of a model. This treatment could be evaluated in a straightforward manner in a new instance of the experiment presented in this paper.

Acknowledgement: We would like to thank all the people who participated in our experiment.

References

[1] Marlon Dumas, Marcello La Rosa, Jan Mendling, and Hajo A Reijers. *Fundamentals of business process management*. Springer, 2013.

[2] Object Management Group. Business Process Model and Notation (BPMN) version 2.0, 2011.

[3] Jan Mendling, Mark Strembeck, and Jan Recker. Factors of process model comprehension—findings from a series of experiments. *Decision Support Systems*, 53(1):195–206, 2012.

[4] Daniel L Moody, Guttorm Sindre, Terje Brasethvik, and Arne Sølvberg. Evaluating the quality of process models: Empirical testing of a quality framework. In *Conceptual Modeling—ER 2002*, pages 380–396. Springer, 2003.

[5] Jan C 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. *Proceedings of the 18th Australasian Conference on Information Systems*, 2007.

[6] Marian Petre. Why looking isn't always seeing: readership skills and graphical programming. *Communications of the ACM*, 38(6):33–44, 1995.

[7] Marcello La Rosa, Arthur HM Ter Hofstede, Petia Wohed, Hajo A Reijers, Jan Mendling, and Wil MP Van der Aalst. Managing process model complexity via concrete syntax modifications. *Industrial Informatics, IEEE Transactions on*, 7(2):255–265, 2011.

[8] Matthias Schrepfer, Johannes Wolf, Jan Mendling, and Hajo A Reijers. The impact of secondary notation on process model understanding. In *The Practice of Enterprise Modeling*, pages 161–175. Springer, 2009.

[9] Hajo A Reijers, Thomas Freytag, Jan Mendling, and Andreas Eckleder. Syntax highlighting in business process models. *Decision Support Systems*, 51(3):339–349, 2011.

[10] Jan Mendling, Hajo A Reijers, and Jorge Cardoso. What makes process models understandable? In *Business Process Management*, pages 48–63. Springer, 2007.

[11] Stefan Zugal, Jakob Pinggera, Hajo Reijers, Manfred Reichert, and Barbara Weber. Making the case for measuring mental effort. In *Proceedings of the Second Edition of the International Workshop on Experiences and Empirical Studies in Software Modelling*, page 6. ACM, 2012.

[12] Jan Recker, Hajo A Reijers, and Sander G van de Wouw. Process model comprehension: The effects of cognitive abilities, learning style, and strategy. *Communications of the Association for Information Systems*, 34(1):9, 2014.

[13] Jan Mendling and Mark Strembeck. Influence factors of understanding business process models. In *Business information systems*, pages 142–153. Springer, 2008.

[14] Joachim Melcher, Jan Mendling, Hajo A Reijers, and Detlef Seese. On measuring the understandability of process models. In *Business Process Management Workshops*, pages 465–476. Springer, 2010.

[15] Andrew Gemino and Yair Wand. A framework for empirical evaluation of conceptual modeling techniques. *Requirements Engineering*, 9(4):248–260, 2004.

[16] Laura Sanchez-Gonzalez, Felix Garcia, Francisco Ruiz, and Jan Mendling. Quality indicators for business process models from a gateway complexity perspective. *Information and Software Technology*, 54(11):1159–1174, 2012.

[17] Kathrin Figl, Jan Mendling, and Mark Strembeck. The influence of notational deficiencies on process model comprehension. *Journal of the Association for Information Systems*, 2012.

[18] Kathrin Figl and Ralf Laue. Cognitive complexity in business process modeling. In *Advanced Information Systems Engineering*, pages 452–466. Springer, 2011.

[19] Volker Gruhn and Ralf Laue. Reducing the cognitive complexity of business process models. In *Cognitive Informatics, 2009. ICCI'09. 8th IEEE International Conference on*, pages 339–345. IEEE, 2009.

[20] GM Muketha, AAA Ghani, MH Selamat, and R Atan. A survey of business process complexity metrics. *Information Technology Journal*, 9(7):1336–1344, 2010.

[21] Laura Sanchez Gonzalez, Felix Garcia Rubio, Francisco Ruiz Gonzalez, and Mario Piattini Velthuis. Measurement in business processes: a systematic review. *Business Process Management Journal*, 16(1):114–134, 2010.

[22] Jorge Cardoso. Evaluating workflows and web process complexity. *Workflow Handbook*, 2005:284–290, 2005.

[23] Elvira Rolon, Jorge Cardoso, Felix Garcia, Francisco Ruiz, and Mario Piattini. Analysis and validation of control-flow complexity measures with bpmn process models. In *Enterprise, Business-Process and Information Systems Modeling*, pages 58–70. Springer, 2009.

[24] Laura Sanchez-Gonzalez, Francisco Ruiz, Felix Garcia, and Jorge Cardoso. Towards thresholds of control flow complexity measures for bpmn models. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, pages 1445–1450. ACM, 2011.

[25] Jan Mendling, Hajo A Reijers, and Wil MP van der Aalst. Seven process modeling guidelines (7pmg). *Information and Software Technology*, 52(2):127–136, 2010.

[26] Jörg Becker, Michael Rosemann, and Christoph von Uthmann. Guidelines of business process modeling. In *Business Process Management*, pages 30–49. Springer, 2000.

[27] Marlon Dumas, Marcello La Rosa, Jan Mendling, Raul Mäesalu, Hajo A Reijers, and Nataliia Semenenko. Understanding business process models: the costs and benefits of structuredness. In *Advanced Information Systems Engineering*, pages 31–46. Springer, 2012.

[28] Ralf Laue and Jan Mendling. Structuredness and its significance for correctness of process models. *Information Systems and E-Business Management*, 8(3):287–307, 2010.

[29] Hajo A Reijers, Jan Mendling, and Remco M Dijkman. Human and automatic modularizations of process models to enhance their comprehension. *Information Systems*, 36(5):881–897, 2011.

[30] Marlon Dumas, Luciano Garcá-Bañuelos, and Artem Polyvyanyy. Unraveling unstructured process models. In *Business Process Modeling Notation*, pages 1–7. Springer, 2011.

[31] Michael Zur Muehlen and Jan Recker. How much language is enough? theoretical and practical use of the business process modeling notation. In *Advanced information systems engineering*, pages 465–479. Springer, 2008.

[32] Jan Recker and Michael Rosemann. Teaching business process modelling: experiences and recommendations. *Communications of the Association for Information Systems*, 25(1):32, 2009.

[33] John Sweller, Paul Ayres, and Slava Kalyuga. *Cognitive load theory*, volume 1. Springer, 2011.

[34] Dennis E Egan and Barry J Schwartz. Chunking in recall of symbolic drawings. *Memory & Cognition*, 7(2):149–158, 1979.

[35] Fred Paas, Juhani E Tuovinen, Huib Tabbers, and Pascal WM Van Gerven. Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist*, 38(1):63–71, 2003.

[36] Daniel L Moody. The "physics" of notations: toward a scientific basis for constructing visual notations in software engineering. *Software Engineering, IEEE Transactions on*, 35(6):756–779, 2009.

[37] Shehnaaz Yusuf, Huzefa Kagdi, and Jonathan I Maletic. Assessing the comprehension of uml class diagrams via eye tracking. In *Program Comprehension, 2007. ICPC'07. 15th IEEE International Conference on*, pages 113–122. IEEE, 2007.

[38] Helen C Purchase, David Carrington, and Jo-Anne Allder. Empirical evaluation of aesthetics-based graph layout. *Empirical Software Engineering*, 7(3):233–255, 2002.

[39] Bonita Sharif and Jonathan I Maletic. An eye tracking study on the effects of layout in understanding the role of design patterns. In *Software Maintenance (ICSM), 2010 IEEE International Conference on*, pages 1–10. IEEE, 2010.

[40] Liat Antwarg, Talia Lavie, Lior Rokach, Bracha Shapira, and Joachim Meyer. Highlighting items as means of adaptive assistance. *Behaviour & Information Technology*, 32(8):761–777, 2013.

[41] Arzu Coltekin, Sara Irina Fabrikant, and Martin Lacayo. Exploring the efficiency of users' visual analytics strategies based on sequence analysis of eye movement recordings. *International Journal of Geographical Information Science*, 24(10):1559–1575, 2010.

[42] Razvan Petrusel and Jan Mendling. Eye-tracking the factors of process model comprehension tasks. In *Advanced Information Systems Engineering*, pages 224–239. Springer, 2013.

[43] Matthias Weidlich, Artem Polyvyanyy, Jan Mendling, and Mathias Weske. Causal behavioural profiles–efficient computation, applications, and evaluation. *Fundamenta Informaticae*, 113(3):399–435, 2011.

[44] Keith D Cooper, Timothy J Harvey, and Ken Kennedy. A simple, fast dominance algorithm. *Software Practice & Experience*, 4(1-10):1–8, 2001.

[45] Gero Decker and Jan Mendling. Process instantiation. *Data & Knowledge Engineering*, 68(9):777–792, 2009.

[46] Wil MP van der Aalst, Kees M van Hee, Arthur HM ter Hofstede, Natalia Sidorova, HMW Verbeek, Marc Voorhoeve, and Moe Thandar Wynn. Soundness of workflow nets: classification, decidability, and analysis. *Formal Aspects of Computing*, 23(3):333–363, 2011.

[47] Mazin Saeed, Faisal Saleh, Sadiq Al-Insaif, and Mohamed El-Attar. Empirical validating the cognitive effectiveness of a new feature diagrams visual syntax. *Information and Software Technology*, 71:1–26, 2016.

[48] Andrew Duchowski. *Eye tracking methodology: Theory and practice*, volume 373. Springer, 2007.

[49] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press, 2011.

[50] Marcel Adam Just and Patricia A Carpenter. Eye fixations and cognitive processes. *Cognitive Psychology*, 8(4):441–480, 1976.

[51] John R Anderson, Dan Bothell, and Scott Douglass. Eye movements do not reflect retrieval processes limits of the eye-mind hypothesis. *Psychological Science*, 15(4):225–231, 2004.

[52] Maarten W Van Someren, Yvonne F Barnard, Jacobijn AC Sandberg, et al. *The think aloud method: A practical guide to modelling cognitive processes*, volume 2. Academic Press London, 1994.

[53] Frank Hogrebe, Nick Gehrke, and Markus Nüttgens. Eye tracking experiments in business process modeling: Agenda setting and proof of concept. In *EMISA*, pages 183–188, 2011.

[54] Zohreh Sharafi, Zéphyrin Soh, and Yann-Gaël Guéhéneuc. A systematic literature review on the usage of eye-tracking in software engineering. *Information and Software Technology*, 67:79–107, 2015.

[55] Natalia Juristo and Ana M Moreno. *Basics of software engineering experimentation*. Springer Publishing Company, Incorporated, 2010.

[56] Andy Field and Graham J Hole. *How to design and report experiments*. Sage, 2002.

[57] Matthias Weidlich, Jan Mendling, and Mathias Weske. Efficient consistency measurement based on behavioral profiles of process models. *Software Engineering, IEEE Transactions on*, 37(3):410–429, 2011.

[58] Jean-Michel Boucheix and Richard K Lowe. An eye tracking comparison of external pointing cues and internal continuous cues in learning with complex animations. *Learning and instruction*, 20(2):123–135, 2010.

[59] Joseph H Goldberg and Xerxes P Kotval. Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, 24(6):631–645, 1999.

[60] Robert JK Jacob and Keith S Karn. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Mind*, 2(3):4, 2003.

[61] Cristina Conati and Christina Merten. Eye-tracking for user modeling in exploratory learning environments: An empirical evaluation. *Knowledge-Based Systems*, 20(6):557–574, 2007.

[62] Jianqiang Shen, Lida Li, Thomas G Dietterich, and Jonathan L Herlocker. A hybrid learning system for recognizing user tasks from desktop activities and email messages. In *Proceedings of the 11th international conference on Intelligent user interfaces*, pages 86–92. ACM, 2006.

[63] Jakob Pinggera, Marco Furtner, Markus Martini, Pierre Sachse, Katharina Reiter, Stefan Zugal, and Barbara Weber. Investigating the process of process modeling with eye movement analysis. In *Business Process Management Workshops*, pages 438–450. Springer, 2013.

[64] Marina Trkman, Jan Mendling, and Marjan Krisper. Using business process models to better understand the dependencies among user stories. *Information and Software Technology*, 71:58–76, 2016.

[65] Zhiqiang Yan, Remco Dijkman, and Paul Grefen. Business process model repositories–framework and survey. *Information and Software Technology*, 54(4):380–395, 2012.

[66] Matthias Kunze, Matthias Weidlich, and Mathias Weske. Querying process models by behavior inclusion. *Software & Systems Modeling*, 14(3):1105–1125, 2015.

[67] Artem Polyvyanyy, Luigi Corno, Raffaele Conforti, Simon Raboczi, Marcello La Rosa, and Giancarlo Fortino. Process querying in apromore. BPM (Demos):105–109, 2015.

[68] Simone Kriglstein, Günter Wallner, and Stefanie Rinderle-Ma. A visualization approach for difference analysis of process models and instance traffic. In *Business Process Management*, pages 219–226. Springer, 2013.

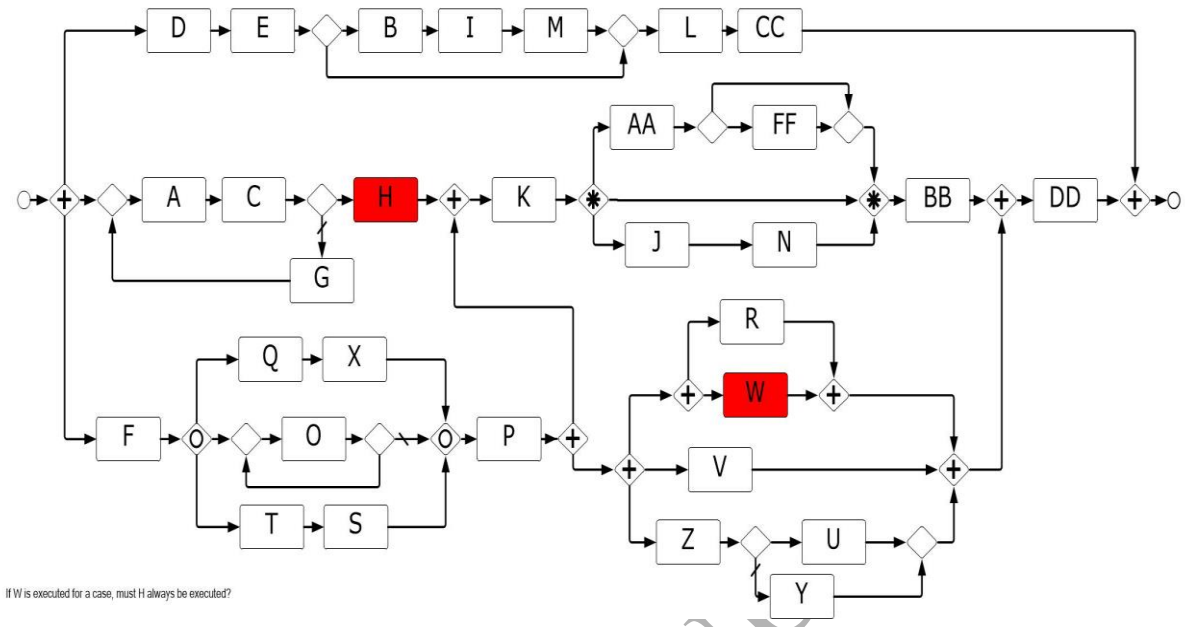
[69] Barbara Weber, Manfred Reichert, Jan Mendling, and Hajo A Reijers. Refactoring large process model repositories. *Computers in Industry*, 62(5):467–486, 2011.

[70] Thomas Gschwind, Jakob Pinggera, Stefan Zugal, Hajo A Reijers, and Barbara Weber. A linear time layout algorithm for business process models. *Journal of Visual Languages & Computing*, 25(2):117–132, 2014.

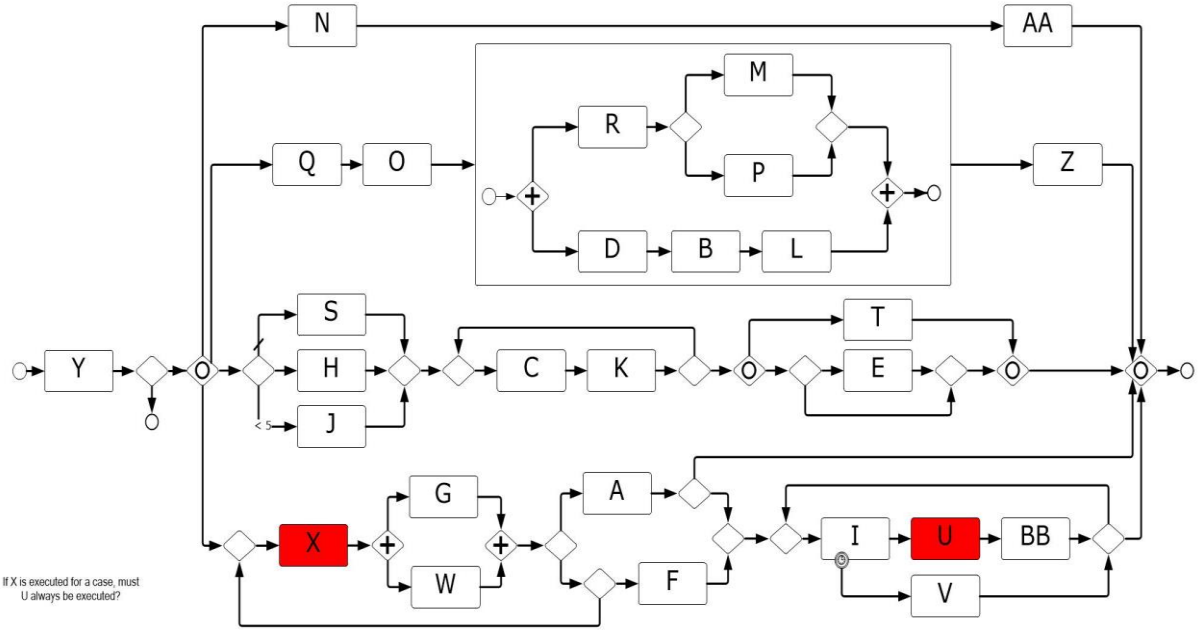
[71] Gove Allen and Jeffrey Parsons. Is query reuse potentially harmful? anchoring and adjustment in adapting existing database queries. *Information Systems Research*, 21(1):56–77, 2010.

[72] Alan Baddeley. *Working memory, thought, and action*, volume 45. OUP Oxford, 2007.

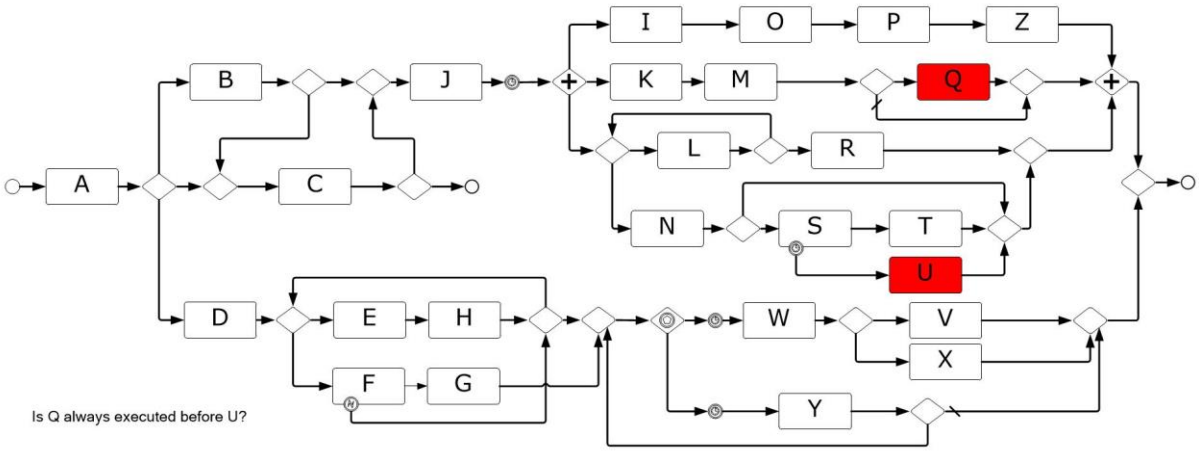
APPENDIX



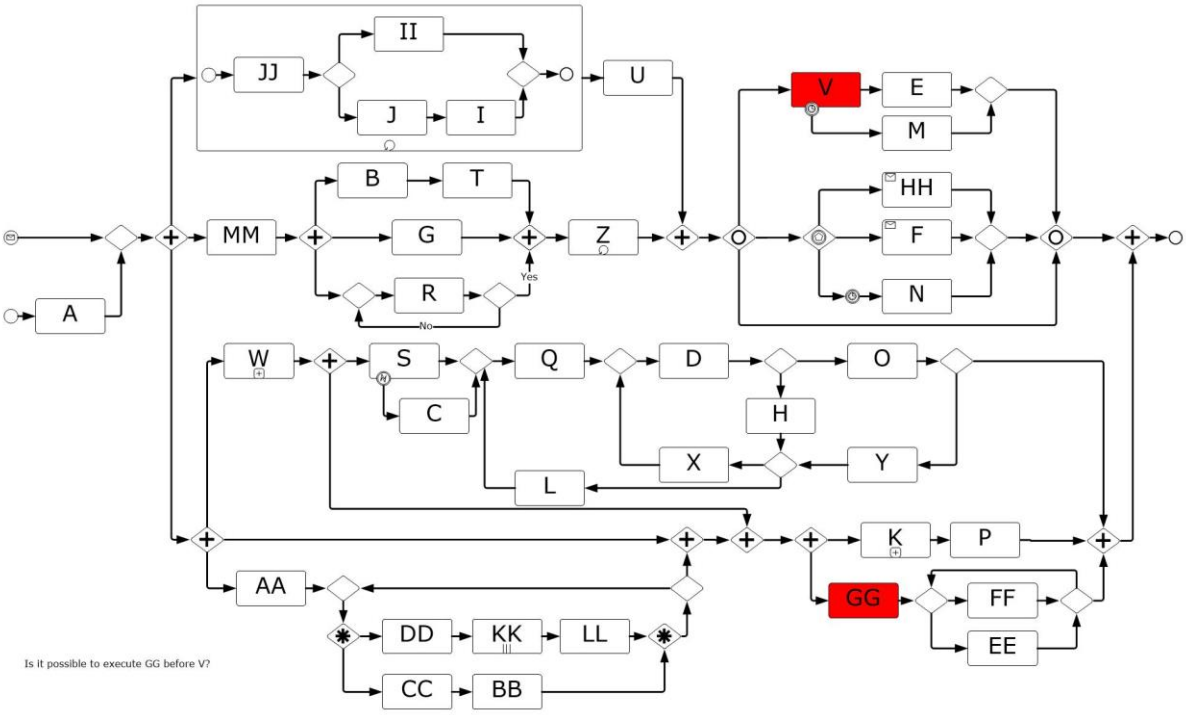
Appendix Figure 1– Benchmark Model 1 and Question 1



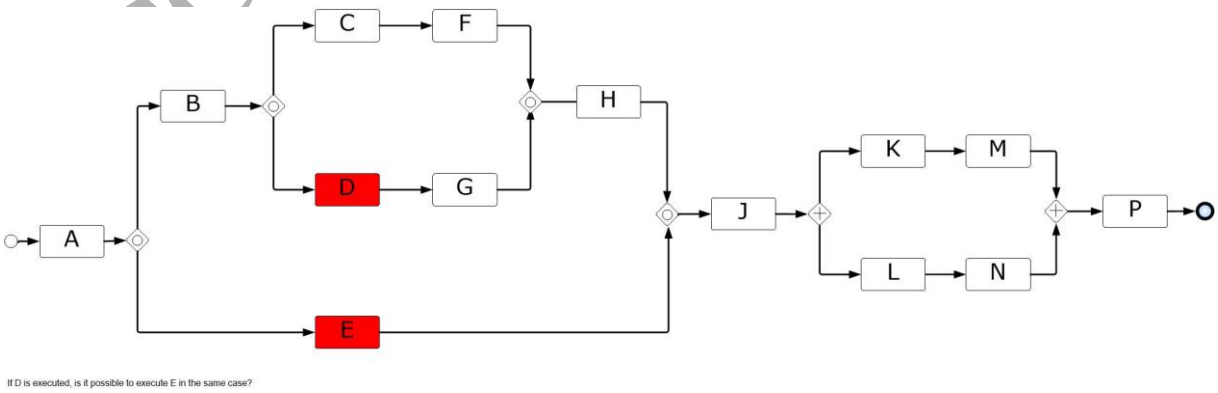
Appendix Figure 2 - Benchmark Model 2 and Question 2



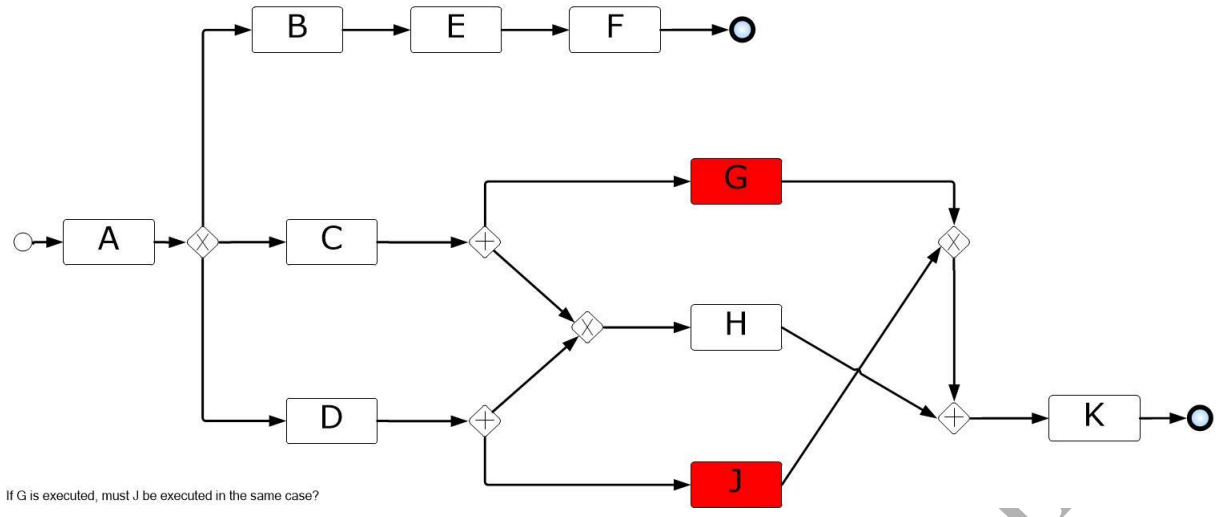
Appendix Figure 3 - Benchmark Model 3 and Question 3



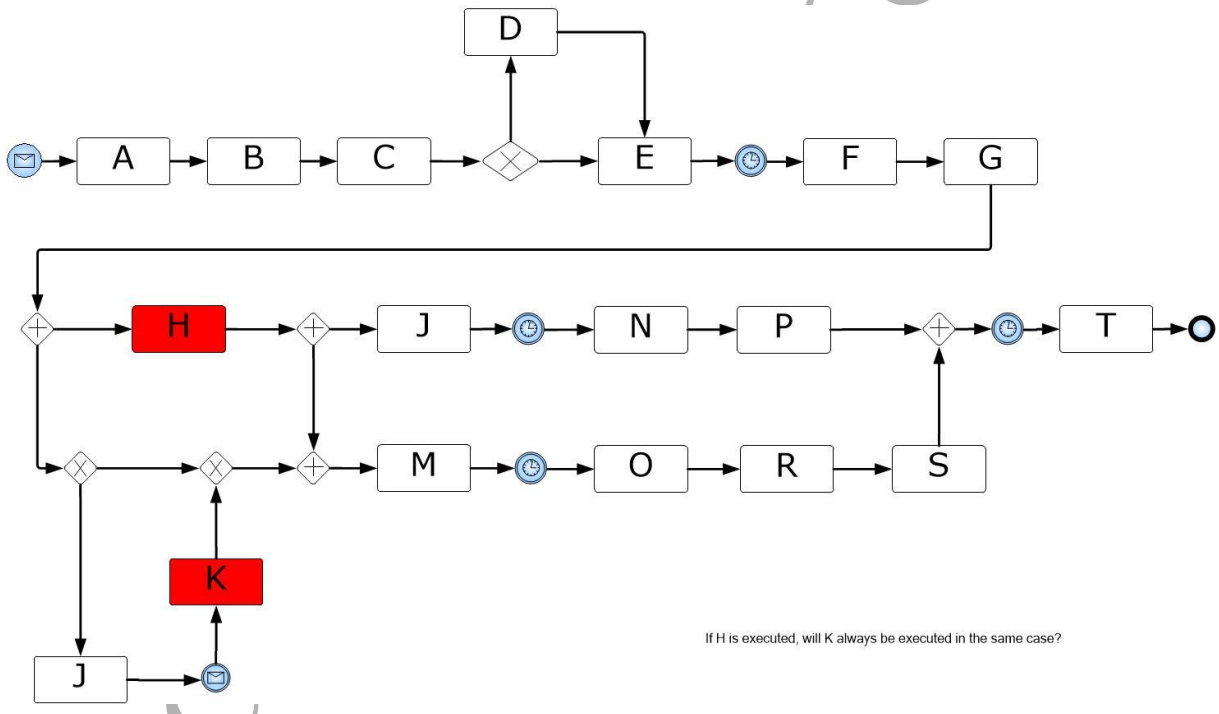
Appendix Figure 4 - Benchmark Model 4 and Question 4



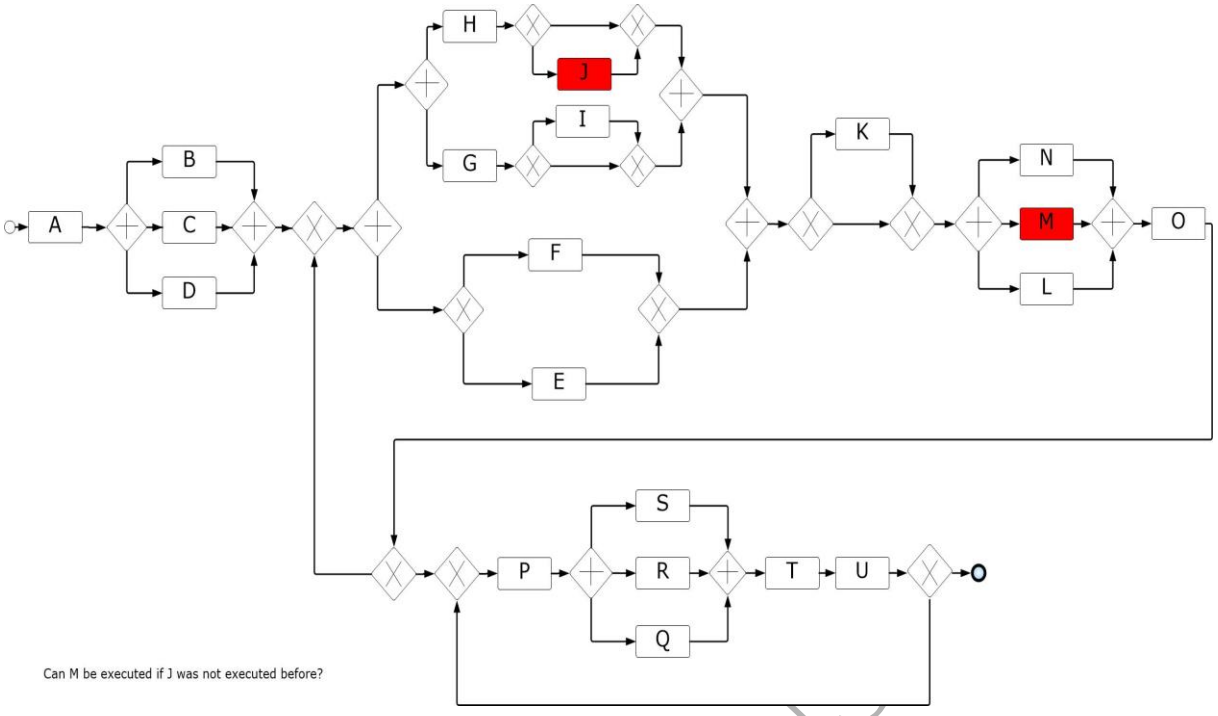
Appendix Figure 5 - Benchmark Model 5 and Question 5



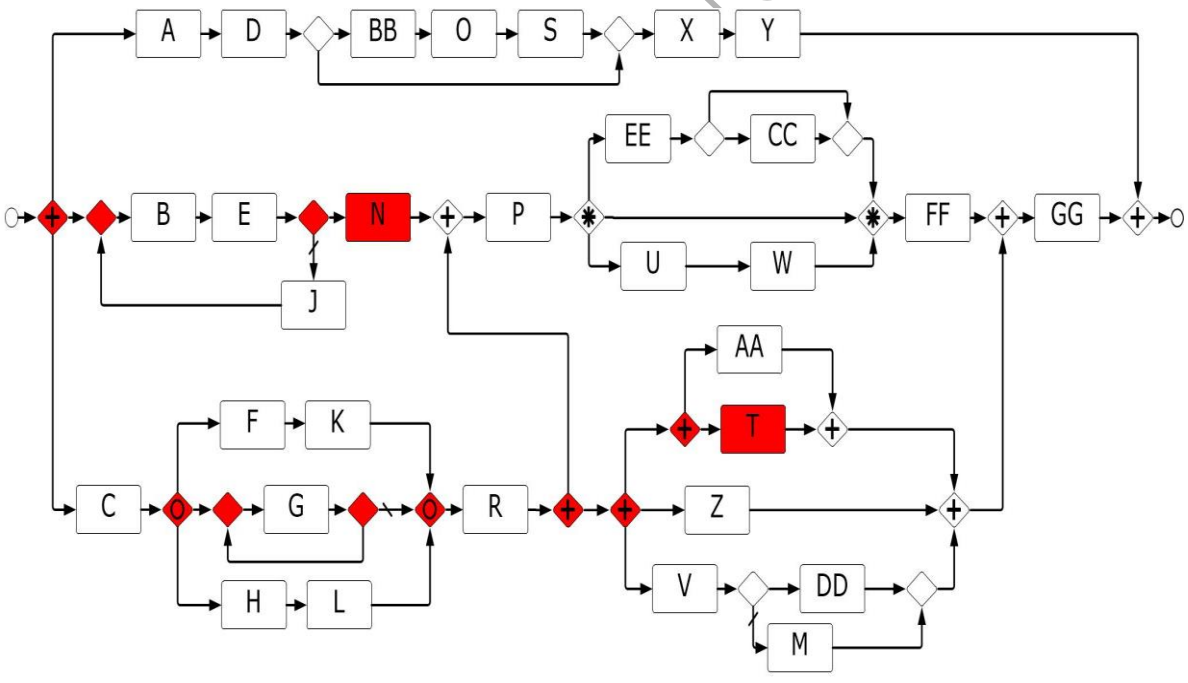
Appendix Figure 6 - Benchmark Model 6 and Question 6



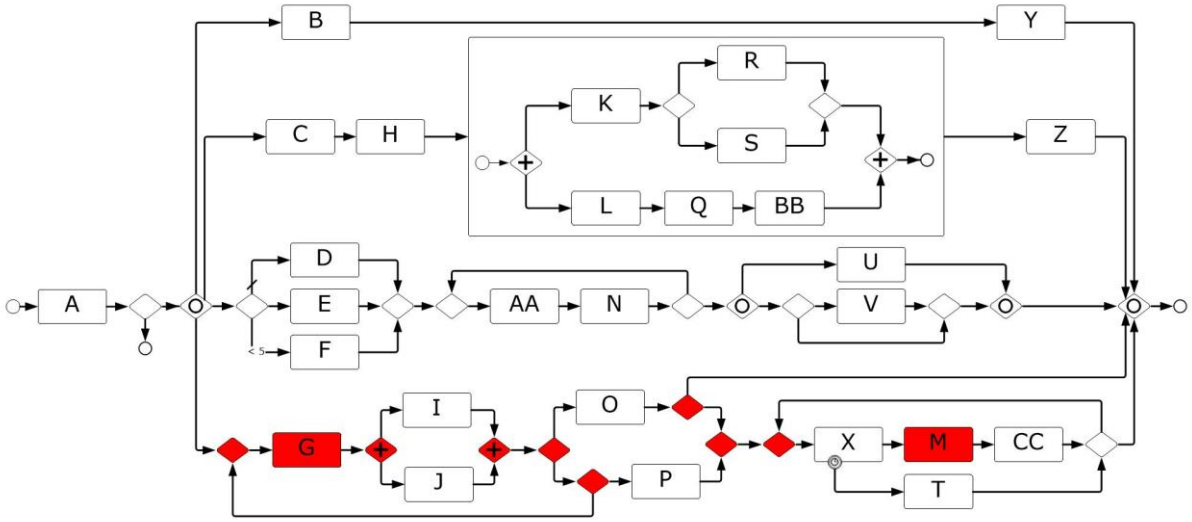
Appendix Figure 7 - Benchmark Model 7 and Question 7



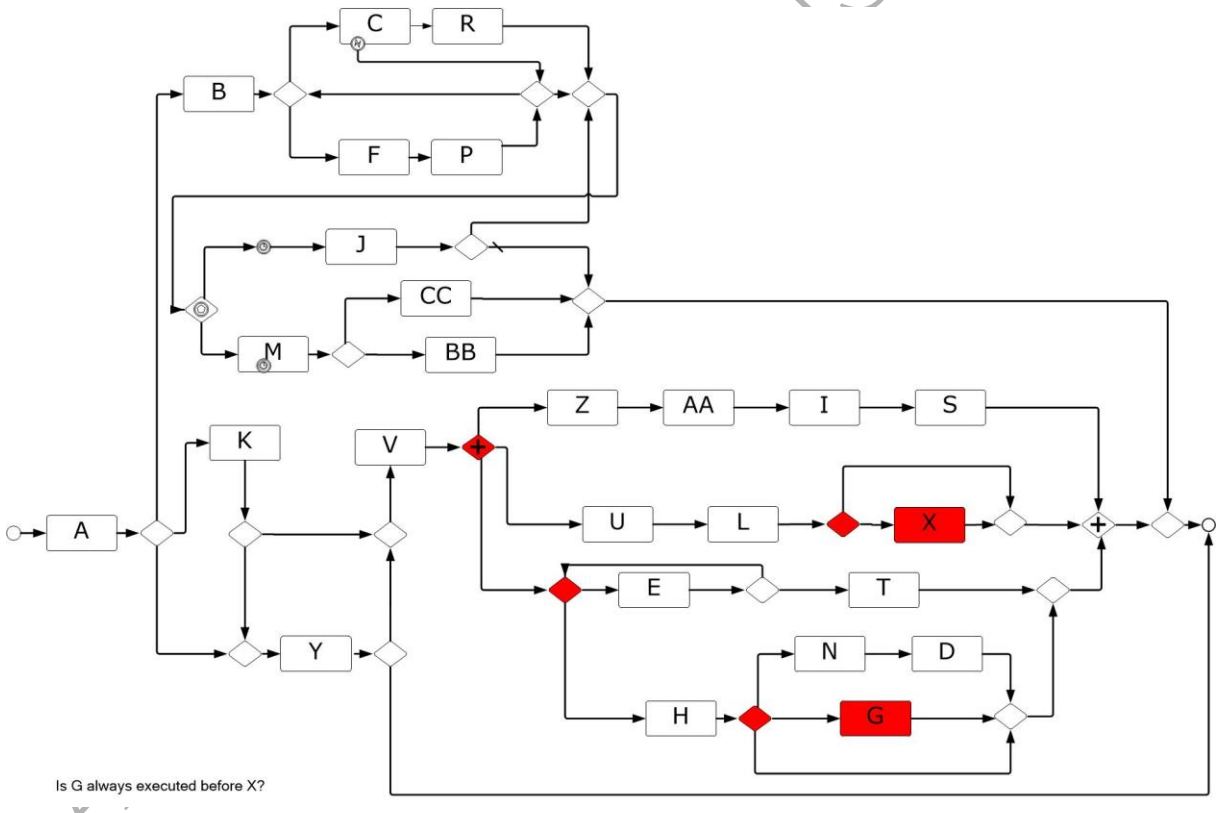
Appendix Figure 8 - Benchmark Model 8 and Question 8



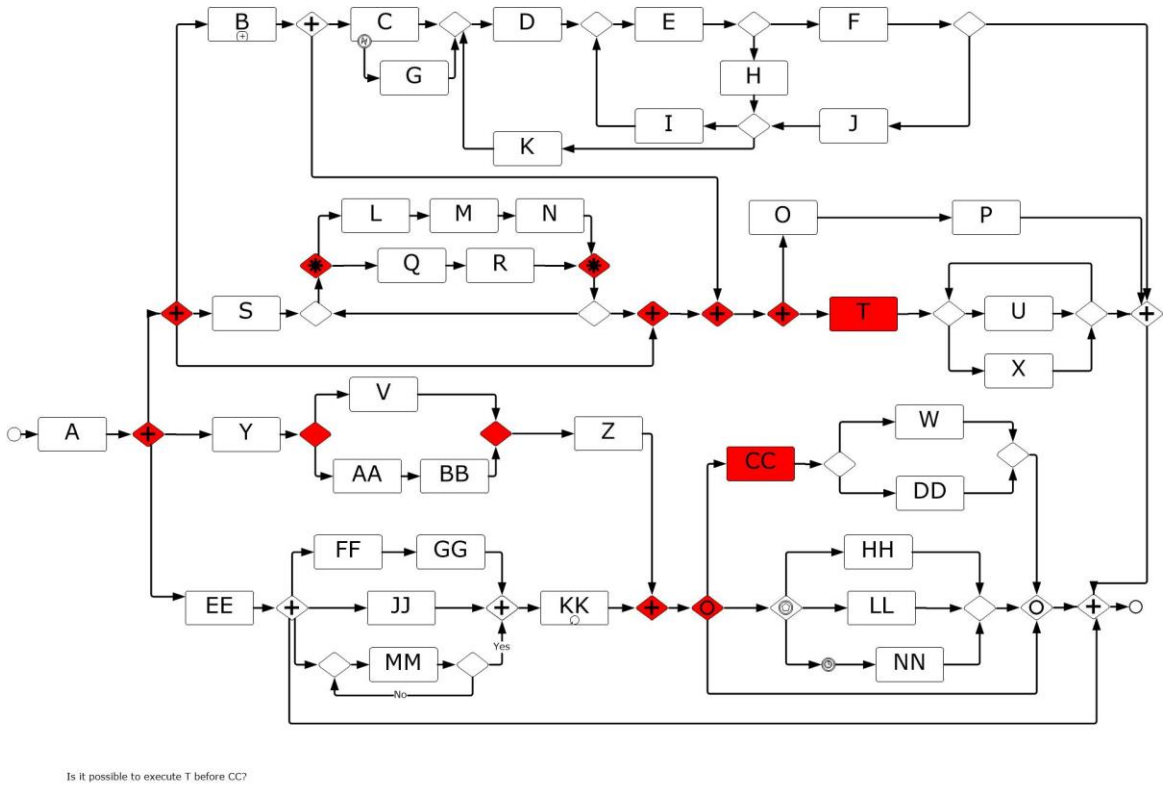
Appendix Figure 9 – Treatment Model 1 and Question 1



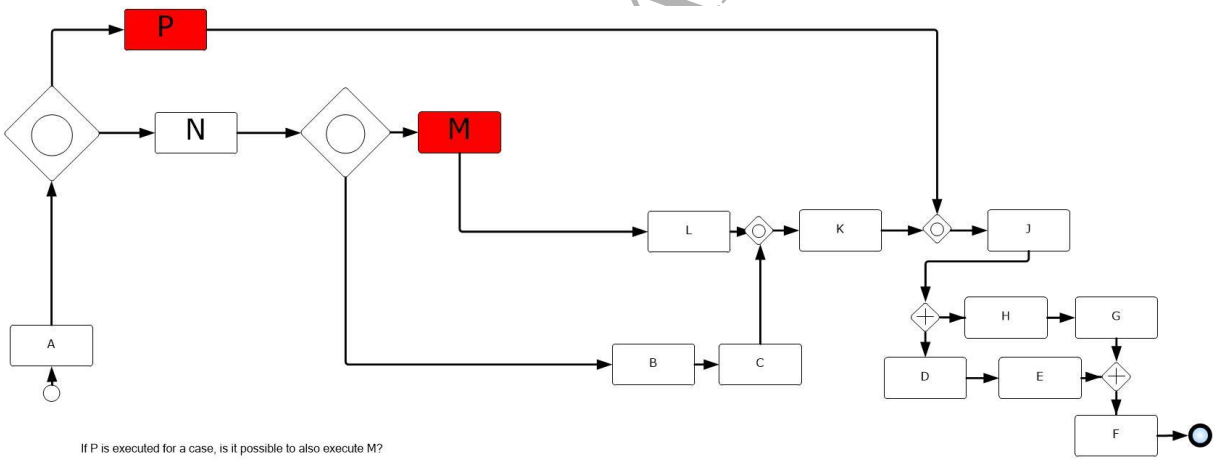
Appendix Figure 10 – Treatment Model 2 and Question 2



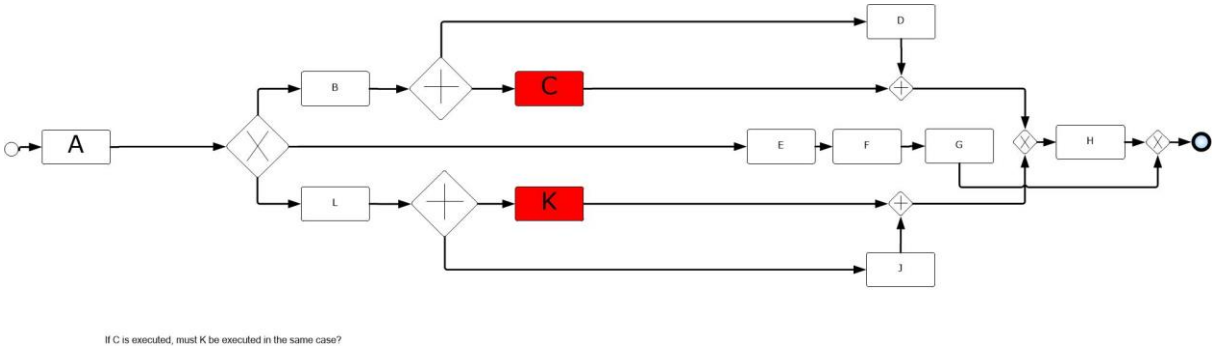
Appendix Figure 11 – Treatment Model 3 and Question 3



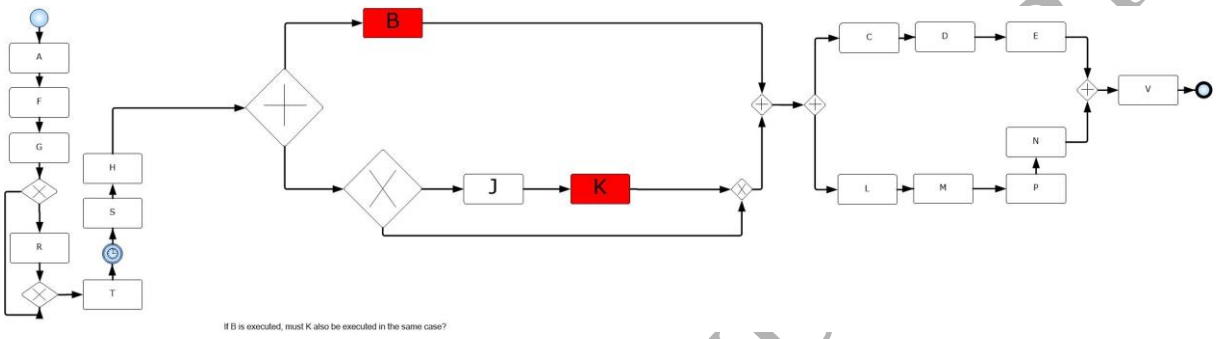
Appendix Figure 12 – Treatment Model 4 and Question 4



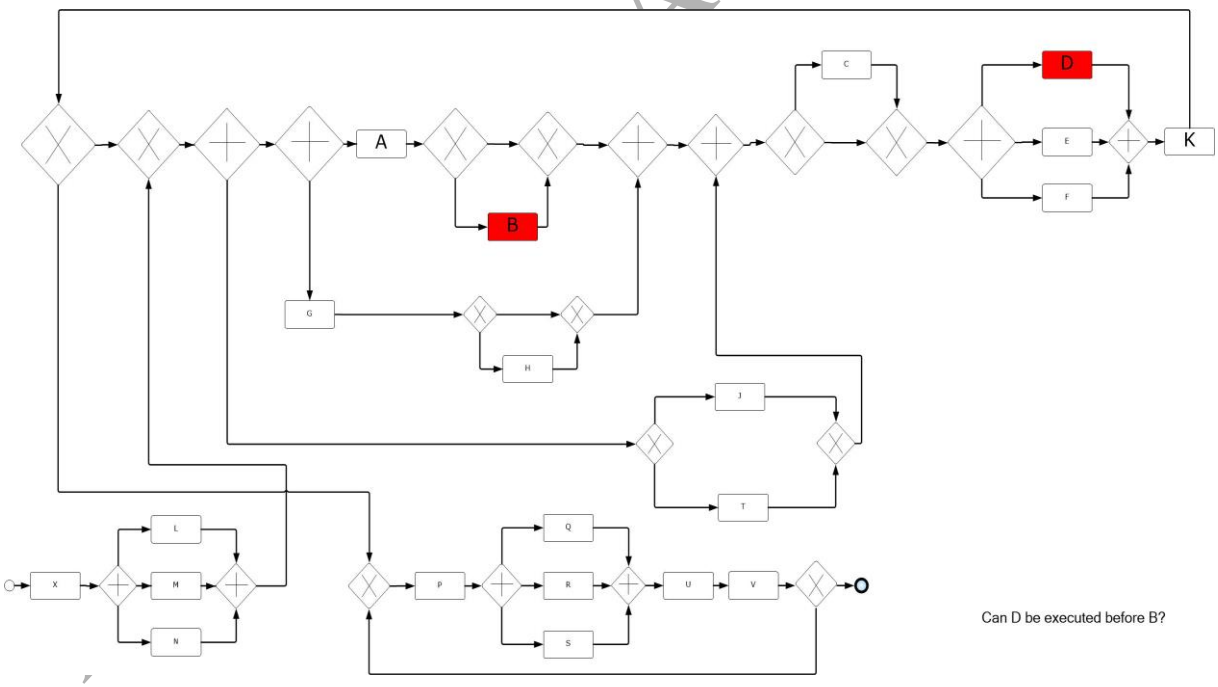
Appendix Figure 13 – Treatment Model 5 and Question 5



Appendix Figure 14 – Treatment Model 6 and Question 6



Appendix Figure 15 – Treatment Model 7 and Question 7



Appendix Figure 16 – Treatment Model 8 and Question 8