UNIVERSITY OF TARTU

FACULTY OF MATHEMATHICS AND COMPUTER SCIENCES

Institute of Computer Sciences

Specialty of Information Technologies

Raul Mäesalu

Complexity and Understandability Comparison between Unstructured and Structured Business Process Models

Master's thesis (30 ECTS)

Supervisors:

Marlon Dumas, PhD

Luciano García-Bañuelos, PhD

Author: …….. "….." May 2011

Supervisor: …….. "….." May 2011

Supervisor: …….. "….." May 2011

Professor:……..…… …… "….." ….. 2011
(name) (signature)

TARTU 2011

Table of Contents

1. INTRODUCTION 3
2. BACKGROUND 6
2.1 UNSTRUCTURED VS. STRUCTURED BUSINESS PROCESS MODELS 6
2.2 UNDERSTANDABILITY AND COMPLEXITY OF BUSINESS PROCESS MODELS 9
3. COMPARATIVE COMPLEXITY STUDY 14
3.1 DESCRIPTION OF DATASET AND MEASUREMENT METHOD 14
3.1.1 IBM dataset 14
3.1.2 Selected metrics 15
3.2 RESULTS 17
3.3 EXAMPLES OF CHANGES IN THE BP MODELS 21
3.4 CONCLUSIONS 27
4. CONTROLLED EXPERIMENT 29
4.1 EXPERIMENTAL DESIGN 29
4.2 RESULTS OF THE EXPERIMENTS 32
4.2.1 Average results 32
4.2.2 Analysis of questions with most incorrect answers 33
4.3 THREATS TO VALIDITY 37
4.4 CONCLUSIONS DRAWN FROM THE EXPERIMENT 39
5. CONCLUSIONS 44
RESÜMEE 47
REFERENCES 48
SUPPLEMENTARY MATERIAL 50

1. Introduction

A business process (BP) is an activity or a set of activities that will accomplish a specific organizational goal [1]. A BP model is a mapping or a visualization of that process using a modeling language. BP modeling has become an important part of the work of many organizations. Using BP models the companies can automate their processes and measure the performance of them in such areas like time and cost. Using these measurements, organizations can find bottlenecks in their processes and they can re-design them to be more efficient, economical and cost-effective. An example of a simple BP model can be seen in Figure 1.

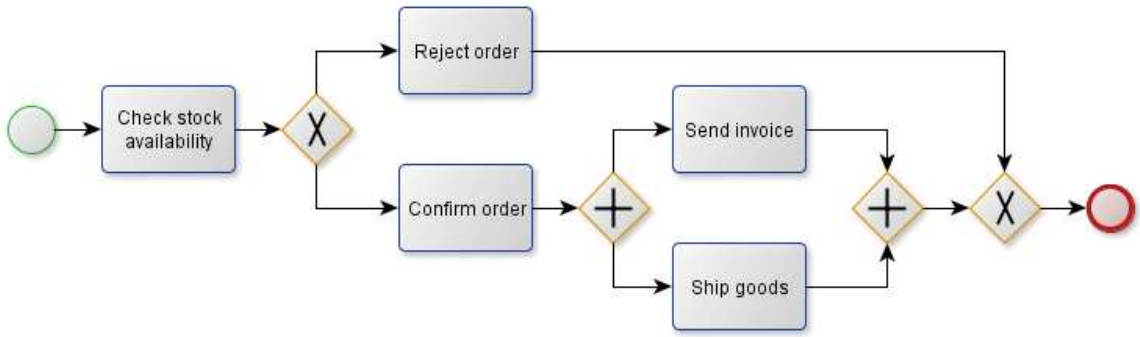


Figure 1: Simple BP model for an order management process.

For any major company, proper and effective use of BP models is essential for its success. BP models have to be syntactically correct and they also have to be easy to comprehend and maintain [2]. A BP model can have any topology. However, it may not always be easy to comprehend the model if it does not follow some structural rules. A very large BP model with overlapping arcs that resemble spaghetti would be a fitting example. Therefore, it is recommended to BP modelers as a guideline that they should try to design their models as structured as possible [3]. A BP model is structured if every split connector matches a respective join connector of the same type. An analogy can be made with formulas with balanced brackets, i.e. every opening bracket has a corresponding closing bracket of the same type. Structured models are less likely to contain errors and are also easier to understand. An example of an unstructured model can be seen on Figure 2(a). The model is unstructured, because the parallel splits *u* and *v* do not have corresponding joins. An equivalent structured model for the given example can be seen on Figure 2(b).

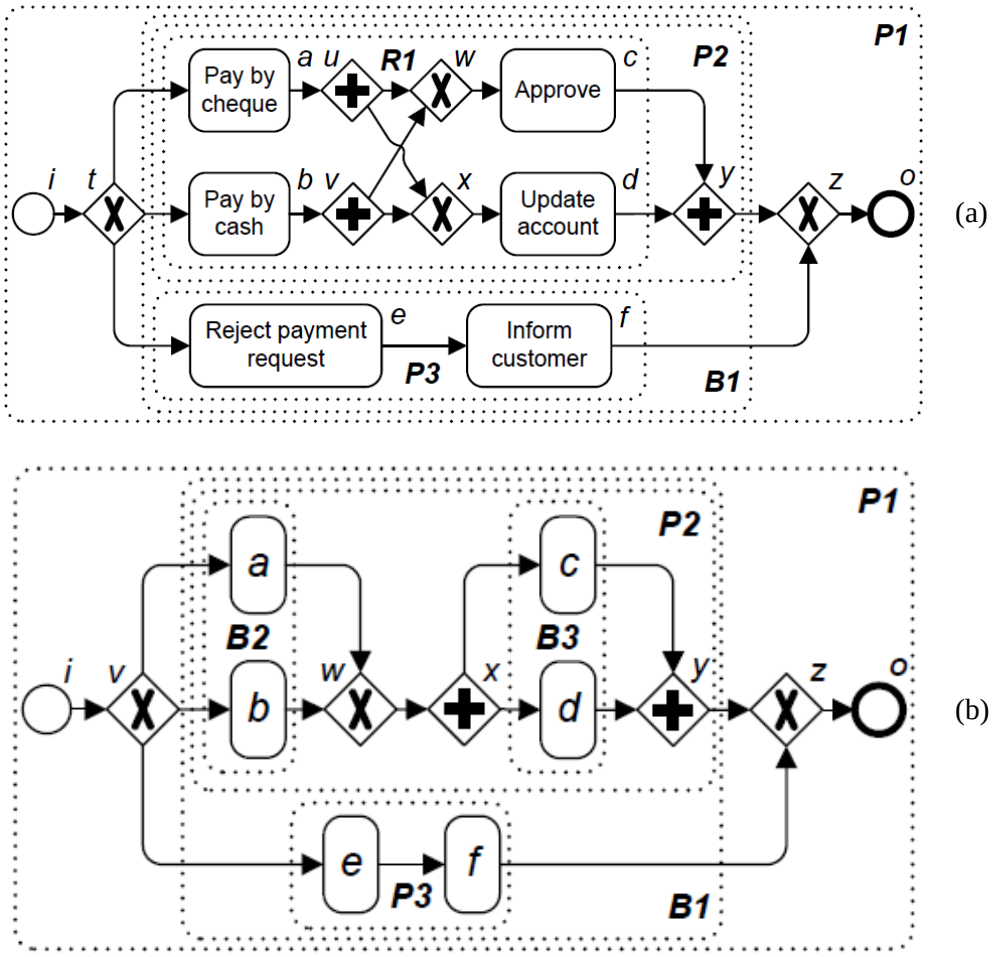


Figure 2: An unstructured BP model (a) and the equivalent structured BP model (b).

In previous academic research, a tool called BPStruct [4] has been developed for automatically converting unstructured BP models into structured ones. Other academic research has suggested that structured BP models are generally less complex and also more understandable. From this we can make the assumption that structured BP models that are produced from unstructured BP models are also easier to comprehend. The purpose of this thesis is to confirm or deny the following two hypotheses:

- H1. BP models restructured with BPStruct are less complex than equivalent unstructured ones;
- H2. BP models restructured with BPStruct are easier to comprehend than equivalent unstructured ones.

The rest of the thesis is structured as follows. Chapter 2 will give a more specific definition to structuredness of BP models, describes the BPStruct tool, reviews related work about BP model complexity and understandability and describes several different metrics that have emerged from academic papers about BP model measurement. Chapter 3 will use a dataset of BP models to generate structured BP models with BPStruct and then measures the models with different metrics that were presented in Chapter 2. The results of the measurements will be analyzed and some example models will be demonstrated. In Chapter 4, an experiment will be conducted to empirically validate the results of the measurements. The results of the experiment will be analyzed and compared to the results from Chapter 3. Finally, Chapter 5 will review the results of this thesis and gives final conclusions to the research hypotheses.

2. Background

This chapter of the thesis will give an overview of the theoretical background of the topic and briefly discusses the related work that has already been done in this area of research. The first section will review the necessary definitions about BP modeling and describe the difference of unstructured and structured BP models. The second section will review the related work done in this research area.

2.1 Unstructured vs. Structured Business Process Models

A formal definition of structuredness states that a well-structured BP model is a model where for every node with multiple outgoing arcs (a split) there is a corresponding node with multiple incoming arcs (a join), such that the set of nodes between the split and the join form a single-entry-single-exit (SESE) region [5]. This section will give a more detailed description of what a well-structured BP model is. Also, a brief overview about how BPStruct works will also be given.

BPStruct uses a technique called Refined Process Structure Tree (RPST) [5] [6] to decompose a BP model into a tree of regions, each representing a SESE region in the model. The root of the RPST represents the entire process model and going down the RPST, the tree consists of smaller and smaller SESE regions until reaching single arcs at the lowest level. The SESE regions (or *components*) of the RPST can be classified into the following four categories:

- A *trivial (T)* component consists of a single edge (e.g. *(i,t)* on Figure 2(a) [7]).
- A *polygon (P)* represents a sequence of components (e.g. *(i,t), B1, (z,o))* on Figure 2(a)).
- A *bond (B)* stands for a set of components that share two common nodes (e.g. {*P2, P3}* on Figure 2(a)).
- A *rigid (R)* is any other component in the BP model that does not fall into any of the three previous categories (e.g. fragment *R1* on Figure 2(a)).

A BP model is structured if its RPST does not contain any rigid components. Figure 2(a) presents an example of RPST decomposition of an unstructured BP model in the form of dotted boxes. R1 is a rigid component. Figure 2(b) presents the RPST of the equivalent structured BP model. It contains only P and B components.

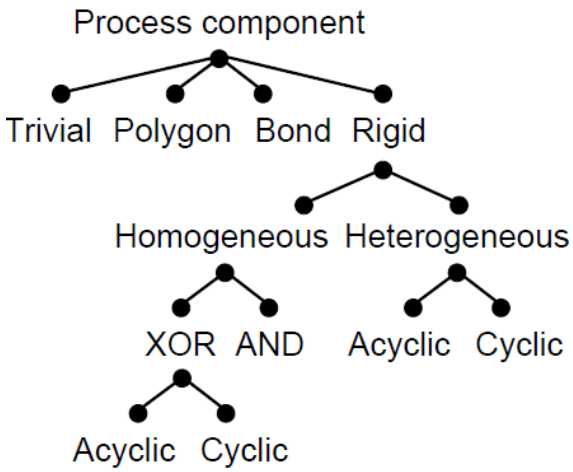


Figure 3: Taxonomy of process components in a BP model.

BPStruct can restructure a BP model if it can restructure every rigid component of that model. Rigid components are classified according to the taxonomy of process components presented in Figure 3. According to this classification of rigid components by types of gateways present in the model, BPStruct will use a different method for restructuring that particular rigid. These methods include Complete Prefix Unfolding, Modular Decomposition Tree and Fully Concurrent Bisimulation. [5] describes these methods and the inner workings of BPStruct in further detail.

However, there are some restrictions for using BPStruct to transform unstructured BP models into structured versions since not all models are structurable. Firstly, only models that are composed of nodes (tasks, events, and gateways) and control flow relations are considered. Models with elements like artifacts, annotations, associations, groups, pools, lanes, message flows, sub-processes and attributes are not supported in BPStruct. Secondly, unsound BP models are also not considered. A process is sound if and only if (a) any case terminates in one of some predefined termination states and (b) for all activities in the process there is at least one case in which they can be executed [8]. Thirdly, only BP models where tasks have only one incoming or one outgoing arc are considered. Finally, OR gateways, complex gateways, error events and non-interrupting events are not handled.

In addition to the above restrictions, some BP models are *inherently* unstructured, which means that they cannot be structured in to an equivalent BP model. To determine, which model is inherently unstructured, BPStruct generates a Modular Decomposition Tree (MDT). In an MDT, there are four types of modules: trivial, linear, complete and primitive modules. A BP model is inherently unstructured if and only if its RPST has a rigid component for which its MDT contains a primitive module [5]. Figure 4 displays an example of an inherently unstructured BP model.

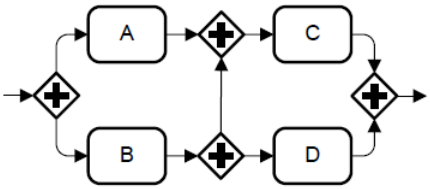
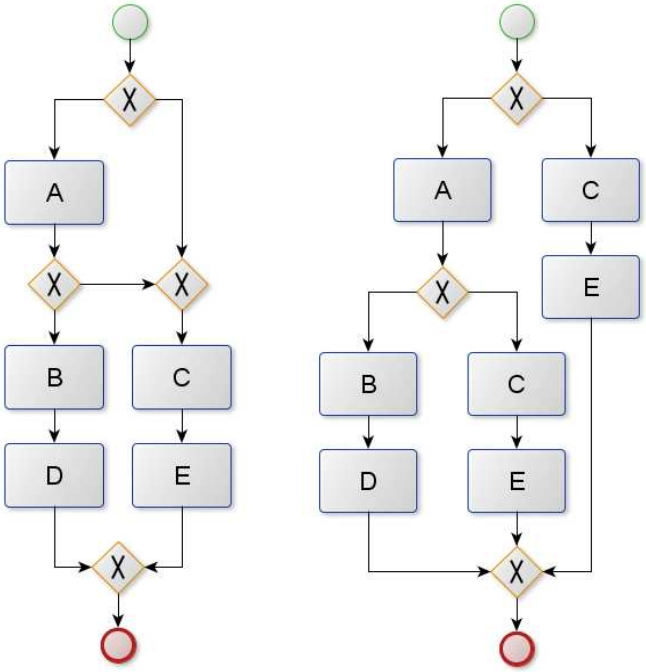


Figure 4: Inherently unstructured BP model.

Notwithstanding these restrictions of BPStruct and the existence of inherently unstructured BP models, there are models that can be transformed into equivalent structured versions. In terms of this thesis, the purpose is to verify whether it is desirable to perform this transformation. Are the BP models restructured with BPStruct less complex and easier to comprehend than their original, unstructured counterparts? One factor that could influence the answers to these questions negatively could be the fact that in order for BPStruct to be able to restructure the BP model, some tasks in it need to be duplicated. This happens, because some edges cannot be drawn freely between nodes in the model. To overcome this restriction, some tasks are duplicated. An example of this task duplication is displayed in Figure 5. In 5(a) an unstructured BP model can be seen. In 5(b) the corresponding structured version can be seen, where tasks C and E have been duplicated.

The reasons why task duplication could influence the complexity and understandability of BP models negatively are simple. Firstly, task duplication makes the models larger in size and larger models could be more difficult to read for human eyes. Secondly, if the same task is located in two different areas of one model, then a person could simply not notice one of them and therefore interpret the BP model wrong. This may generate confusion and increase error-proneness. However, the resulting models that are well-structured have also increased modularity. The hypothesis is that low modularity generally relates to more

errors than higher modularity [9]. This gives grounds to presume that they are also easier to understand. Therefore, the answers to the research questions presented in this thesis lie on the balance between the effect of greater modularity that structured BP models bring, and the effects of having duplicate tasks in the structured BP model.



(a) (b)

Figure 5: Example of task duplication caused by structuring a BP model. (a) is the unstructured model. (b) is the equivalent structured model with tasks C and E duplicated.

In the next section of this chapter an overview is given about related work in the area of measuring complexity and understandability of BP models. A set of metrics will be identified that have been mentioned in several research papers as significant indicators of complexity and understandability. From these metrics, a selection will be made in order to measure models before and after restructuring them with BPStruct.

2.2 Understandability and Complexity of Business Process Models

There has not been much work done in the area of measuring complexity and understandability of BP models, which is evident from the small amount of papers available about the topic. However, within the published papers that are available, some common metrics and approaches have emerged. There are several complexity metrics that

have recurred in different research papers done in this area. A couple of these metrics, like average connector degree and density, have also been identified as having a correlation to perceived understandability of BP models [10], which make them suitable candidates for analysis done in this thesis.

Size is a common metric that has been cited in [10] and [11] as being a metric that has been empirically validated as an indicator of model complexity. Evidence has been provided that larger real-world BP models tend to have more formal flaws than smaller models. It is hypothesized that humans who do the modeling lose track of interrelations in large and complex models due to their limited cognitive abilities. In BP models, size can be measured by simply counting all elements within the model like tasks (including start and end events) and gateways. As an alternative, these elements can be measured separately. If necessary, the number flow elements (arcs) may be counted as well. In the example given in Figure 2, size of model in 2(a) is 15 and model in 2(b) is 13.

Both [9] and [12] describe the strong analogy between the domains of software engineering and BP modeling. The analogy can be demonstrated by similarities between software programs constructs and business processes. Two sequential software statements can be mapped to two sequential process activities. A ‘switch’ statement can be mapped to an XOR-split, threads can be mapped to AND-splits and ‘if-then’ statements can be mapped to OR-splits. [9] uses five design principles of software engineering (coupling, cohesion, complexity, modularity, size) and describes metrics for BP modeling using the same principles. [12] presents some of the more popular metrics in software engineering and tries to adapt them to BP modeling: Lines of Code, McCabe’s Cyclomatic Complexity, Halstead Complexity metric, etc. Common metrics of these two papers include different measurements of size and the Control-flow complexity (CFC) metric, which is based on McCabe’s Cyclomatic Complexity.

The CFC metric evaluates the complexity introduced in a process by the presence of XOR-split, OR-split, and AND-split constructs. In this thesis, OR-splits are not considered and are not necessary to measure. For XOR-splits, the CFC is the fan-out of the split. For an AND-split the complexity is simply 1. Mathematically, CFC is additive, which means that to get the CFC of a BP model, the CFC of all split constructs needs to be added together [9]. The value of CFC should correspond to the values of McCabe’s Cyclomatic

Complexity for which in practice, the industry interpretation is the following: from 1 to 10, the program (in our case the model) is simple; from 11 to 20, it is slightly complex; from 21 to 50 it is complex; and above 50 it is untestable. In the paper that first introduced the metric [13], a small experiment was conducted, which demonstrated a correlation between perceived complexity and CFC. In the example provided in Figure 2, CFC for 2(a) is 5 and CFC for 2(b) is 4. According to these values both models are simple, but example 2(b) is slightly less complex.

The Cross-connectivity (CC) metric that is described in [14] was specifically designed to add to the lacking amount of metrics for the research area of BP model measurement. A study was conducted to validate it in terms of error prediction and understandability. For the former, the study confirmed the hypothesis that it indeed does have a correlation to error probability in BP models. For the latter, the authors concluded that there is a relation between CC and perceived understandability, but it is less powerful than the two best candidate metrics available, which are average connector degree and density. The metric expresses how tightly the nodes in a process model are connected building on a weakest-link metaphor. It also considers all nodes as unique, even if their (business) semantics may be the same; this means that it supports duplicate tasks. CC is calculated in a way, where all nodes of the BP model get a weight value. A lower value is given to connectors that have a higher degree, i.e. they have more options in choosing the path that is taken. The values of nodes are used to calculate all paths between the nodes and divided by the number of total nodes times the number of total nodes minus one. The definition of the metric builds on the assumption that a higher value is associated with an easier understanding of the model, which implies as a consequence a lower error-probability. In the example provided in Figure 2, CC for 2(a) is 0,07541 and CC for 2(b) is 0,08907. According to this value, example 2(b) is easier to understand and also less error-prone.

In [10], a survey was conducted among students of three European universities to identify metrics that are in relation with perceived understandability of a BP model. Five metrics showed a significant correlation: number of joins, density, average connector degree, potential routing elements mismatch and connector heterogeneity. Out of these five, two metrics, namely density and average connector degree, were the ones that most convincingly related to model understandability and will be described in further detail below. As for the other three, number of joins counts together the amount of joins in a BP

model, mismatch is calculated on the basis of degree and summed up per routing element and connector heterogeneity implies which types of routing elements appear in the model.

Density metric relates the number of available connections to the number of maximum connections for the given number of nodes. The simplest model would be a perfectly sequential model that would have 0 as its density. The most complex model would have an arc between every node in that model and have density as 1. In further detail, and how to exactly calculate it, density is described in [15], where it is also confirmed by an empirical study that it can be used successfully for error prediction, however the authors state that there is room for improvement. In the example provided in Figure 2, density for 2(a) is 0,1818 and density for 2(b) is 0,1875. In this case, density shows that 2(b) is slightly more complex than 2(a).

Average connector degree (ACD), which is also called coefficient of connectivity, refers to the average number of connections that a node has with other nodes of the BP model [9]. Considering the syntax of a BP model, then the minimal ACD in a correct model would be 3. Higher values would mean that one connector splits the flow into more arcs than the minimum, which intuitively would make the model more difficult to understand. In the example provided in Figure 2, ACD for 2(a) is 3,1429 and ACD for 2(b) is 3,2. This would mean that model 2(b) is slightly more complex and more difficult to understand than 2(a).

A survey for empirical validation of perceived understandability is also done in [11], where three categories of factors that potentially influence it are identified. They are personal, structural and textual factors. Metrics that had the most significant correlation to perceived understandability were theoretical knowledge, separability of the model and textual length of task labels. Comparing meaningful labels to abstract labels did not demonstrate a significant difference. Figure 2 illustrates meaningful and abstract label examples, where 2(a) has meaningful task labels and 2(b) has abstract task labels. Theoretical knowledge metric was acquired by having the participants of the survey answer six theoretical yes/no questions about BP modeling. Separability is a metric which relates to the number of nodes in a model whose deletion separates the model into multiple components. In Figure 2, the separability of both models is 2.

The next chapter of the thesis will provide a selection of metrics discussed above. That selection of metrics will then be used to perform a comparative complexity study for BP models before and after restructuring them with BPStruct.

3. Comparative complexity study

This chapter summarizes the results of a quantitative analysis aimed at comparing the complexity of unstructured BP models and their corresponding structured versions on the basis of a subset of the complexity metric introduced in the previous chapter. Firstly in this chapter, a dataset of BP models will be introduced for which the study about complexity and understandability will be based on. Secondly, a set of complexity metrics that were discussed in the previous chapter are selected. Thirdly, the data will be measured with named metrics and results of the measurements will be presented and then analyzed. Finally, conclusions will be drawn from the results of these measurements. The next section will concentrate on introducing the dataset and the selection of complexity metrics.

3.1 Description of dataset and measurement method

3.1.1 IBM dataset

To restructure BP models and gather data about them, the dataset of IBM WebSphere Business Modeler process models (IBM dataset) [16] was selected. The IBM dataset was specifically created for [17] in order to analyze methods of checking BP model soundness. Originally, it contained 735 different models represented in an IBM proprietary file format used by IBM WebSphere that contained combined elements from UML Activity Diagrams and Business Process Modeling Notation (BPMN) [18]. The dataset was cleansed by the authors of BPStruct and stored in Event-Driven Process Chains (EPC) Markup Language (EPML) format due to its simplicity compared to other formats. The cleansed IBM dataset contained a total of 533 BP models in EPML. The EPCs were not in a valid format, as they only contained functions and no events. A correct EPC contains both events and functions [19]. However, for the purposes for this thesis, the validity of the EPCs was not important, as the BP models were considered as analogous to BPMN containing only tasks and gateways. This was actually beneficial, since it eliminated one of the BPStruct restrictions introduced in Chapter 2, where elements like artifacts, groups, pools, lanes, etc. were not supported.

The dataset was divided into three anonymised libraries called A, B3 and C. There were 269, 247 and 17 models in each of them accordingly. The models themselves were also in an anonymous format as the label names were abstract (e.g. s00000982, s00001088, etc.) and they only represent purely structural information. The models originate from the real-life domains of insurance, banking, customer relationship, construction and automotive supply chain. The IBM dataset is suitable for the purposes of this thesis, because it will show whether BP models used in real life can be restructured and whether the outputs are in a simpler and more understandable format.

Not all of the 533 BP models were in accordance with the restrictions of BPStruct. There were models that were already structured, contained OR-gateways and were unsound or inherently unstructurable. Every model that created a conflict with BPStruct was removed from the dataset. In the end, there were 59 structurable models for which BPStruct could generate an output. 41 of them were in library A, 15 in library B3 and 3 in library C. For an overview of the IBM dataset, see Table 1.

Table 1: Overview of IBM dataset.

Library	Models initially	Restructurable models
A	269	41
B3	247	15
C	17	3
Total	**533**	**59**

3.1.2 Selected metrics

Resulting from the analysis made on the available metrics in the BP modeling domain, the following metrics were chosen for the purposes of this thesis:

- Number of arcs (#arcs)
- Number of gateways (#GW)
- Number of tasks (#tasks)
- Size of the model (size)
- Control-flow complexity (CFC)
- Cross-connectivity (CC)

- Average connector degree (ACD)
- Density

Size was selected because it is a metric for which it has been empirically validated that a larger BP model is generally more error-prone, and therefore should be more difficult to comprehend. Related metrics, #arcs, #GW, #tasks (includes start and end events) provide a good comparison point for both size and other metrics. With #GW, it was first considered to analyze the number of XOR-gateways and the number of AND-gateways separately, but since AND-gateways were present in only 18,6% of the models then their values did not offer enough meaning. Instead, it was opted to use the total number of gateways. CFC was selected because it is a complexity metric that was recurrently brought up in several research papers about BP modeling metrics. CC has demonstrated in previous studies that it can be related to both error prediction and model understandability. ACD and density were chosen, because in academic literature, they are referred to as the two best metrics available to measure model understandability.

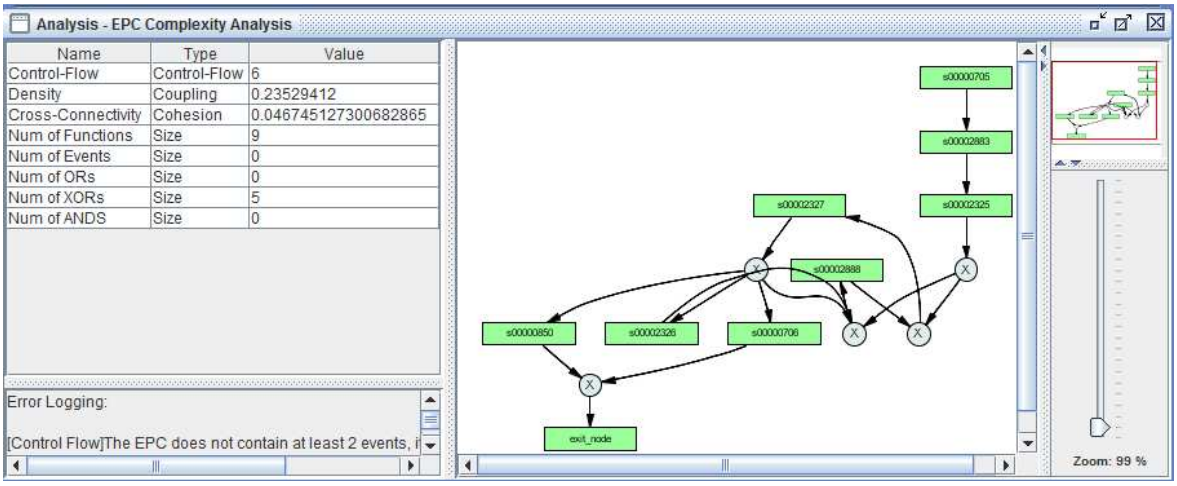


Figure 6: EPC Complexity Analysis plug-in in ProM.

In order to calculate the values of mentioned metrics, two main approaches were taken. Firstly, an open-source tool called ProM 5 [20] was used. It is a process mining framework that supports a wide variety of techniques for mining processes, implemented as plug-ins. For the purposes of this thesis, ProM was useful, because it supports importing EPCs in EPML format and it also contains a plug-in called EPC Complexity Analysis to measure their complexity. An example of this ProM plug-in can be seen in Figure 6. #GW, #tasks, Size, CFC, CC and Density were measured using this approach. Using ProM, every model

had to be imported and analyzed one-by-one. Secondly, measuring #arcs and ACD was implemented as a UNIX shell script, using mainly the grep tool [21] to handle the EPML files and to find and count the necessary elements involved. The collected data was stored and handled in a Microsoft Office Excel spreadsheet. The spreadsheet is given as a supplementary material to the thesis in the form of a companion CD.

The next section of this thesis will give an overview of measurement results.

3.2 Results

In this section of the chapter, an overview will be given about the gathered measurements.

We begin by examining the average values of measured metrics of the BP models before and after restructuring them with BPStruct. On average, the size of the model increased 46,53%. As explained in Section 2.1 of the thesis, some edges cannot be drawn freely in the restructured model and tasks have to be duplicated to overcome this restriction. This explains the general increase in size, because both #tasks and also #arcs are increased this way. On average, #tasks were increased by 53,23% and #arcs were increased by 49,12%. In addition, also #GW were increased by 46,37%. The definition of well-structuredness of a BP model states that for every split there is a corresponding join, such that the set of nodes between them form a SESE region. In case of unstructured models, many splits do not have this corresponding join. BPStruct corrects this by adding the necessary joins to the models. This explains the large increase of #GW in the measurements. For a full overview of the average values of all measured metrics, please refer to Table 2.

The average CFC of the models increased by 35,25% .To explain this increase, we have to look at the way CFC is measured and also remind the working principles of BPStruct. CFC is an additive metric that sums together the fan-out of XOR-splits and BPStruct duplicates tasks of the model when restructuring it, which means that in case of some splits, there will be more arcs leading out of them and into the created duplicate task. Adding these reasons together explains the increase of the metric. According to the classification of the CFC metric into simple, slightly complex, complex and untestable models, we see that the

amount of simple models decreases and the amount of more complex models increases. The exact numbers are presented in Table 3.

Table 2: Average values of each metric before and after restructuring the models.

Metric	Average before restructuring	Average after restructuring	Increase/decrease (%)
#arcs	30,02	44,76	49,12%
#GW	8,41	12,31	46,37%
#tasks	16,81	25,76	53,23%
Size	25,22	38,07	46,53%
CFC	9,97	14,27	35,35%
CC	0,05118	0,04455	-11,44%
ACD	3,41590	3,29957	-3,04%
Density	0,16096	0,12376	-22,5%

Table 3: Number of models by CFC classification

Class	Amount before restructuring	Amount after restructuring	Increase/decrease (%)
Simple (CFC 0-10)	39	28	-28,21%
Slightly complex (CFC 11-20)	18	24	33,33%
Complex (CFC 21-50)	2	5	150,00%
Untestable (CFC 50-…)	0	2	200,00%

In case of the CC metric, the average value dropped 11,44% after the restructuring of the models. As stated in Chapter 2, CC is calculated in a way where every node gets a weight based on the amount of choices that can be taken at the node. More choices results in a lower weight for that node. According to those weights, all paths between nodes are calculated and divided by the amount of total nodes times the amount of total nodes minus one. Considering this, the fact that the restructured models have both more nodes due to task duplication and more paths to take to reach the duplicated tasks due to the added

XOR-gateways, results in the fact that the average value of CC is smaller in the restructured versions of the measured models.

The average value of the ACD metric has decreased by 3,04% in the restructured versions of the models. This can be explained by the fact that the new versions of the models have a slightly lower gateway to tasks ratio than the original ones. However, this value is quite low due to the fact that the models have a low ACD to begin with comparing to the minimum value of what ACD can have in a correct BP model. The average value of the density metric has decreased 22,5% in the structured versions of the models comparing to the original ones. Since density relates the number of available connections to the number of maximum connections then it is calculated similarly to the ACD metric then the simultaneous decrease of density is normal.

As the next step, let us take a look at how the metrics are related to one another. To do that, Pearson correlation will be used. By definition, Pearson correlation measures the degree and direction of linear relationship between two variables [22]. The possible values of it are between -1 and 1, where the latter displays a perfect positive correlation and the former displays a perfect negative correlation. A value of 0 displays the lack of correlation between the two variables. The calculations of the correlations of the metrics in this thesis are based on the difference between the metric value in the structured and the metric value in the original model. To calculate the statistical significance of the Pearson correlations, student's t-test was used. A value under 5% is considered statistically significant.

All size metrics are very strongly correlated to one another. However, this is an expected indicator, bearing in mind that BPStruct adds both gateways and tasks to the models in the restructuring process. Also, the logical conclusion is that if nodes are added to the model then arcs are also added. The highest correlation between size metrics is between #arcs and size, which has a perfect positive relation. #tasks and #GW have the lowest value 0,94, which is also a very strong relation. All other correlations between different size metrics fall between these two values. The statistical significance of the size metrics is below 0,38% in every case showing that the correlations are statistically significant. Exact numbers of all calculated correlations can be found in Table 4.

Table 4: Pearson correlation values between all measured metrics and their statistical significance values (in brackets).

	#arcs	#GW	#tasks	Size	CFC	CC	ACD
#GW	0,98 (0,009%)						
#tasks	0,99 (0,14%)	0,94 (0,0006%)					
Size	1 (0,38%)	0,97 (0,003%)	0,99 (0,09%)				
CFC	0,98 (0,007%)	0,98 (18,0%)	0,95 (0,0002%)	0,97 (0,002%)			
CC	-0,31 (0,02%)	-0,31 (0,09%)	-0,31 (0,003%)	-0,31 (0,009%)	-0,31 (0,12%)		
ACD	-0,06 (0,02%)	-0,2 (0,07%)	0 (0,002%)	-0,07 (0,008%)	-0,07 (0,09%)	0,31 (0,41%)	
Density	-0,49 (0,02%)	-0,56 (0,08%)	-0,46 (0,003%)	-0,5 (0,009%)	-0,5 (0,11%)	0,52 (0%)	0,71 (2,49%)

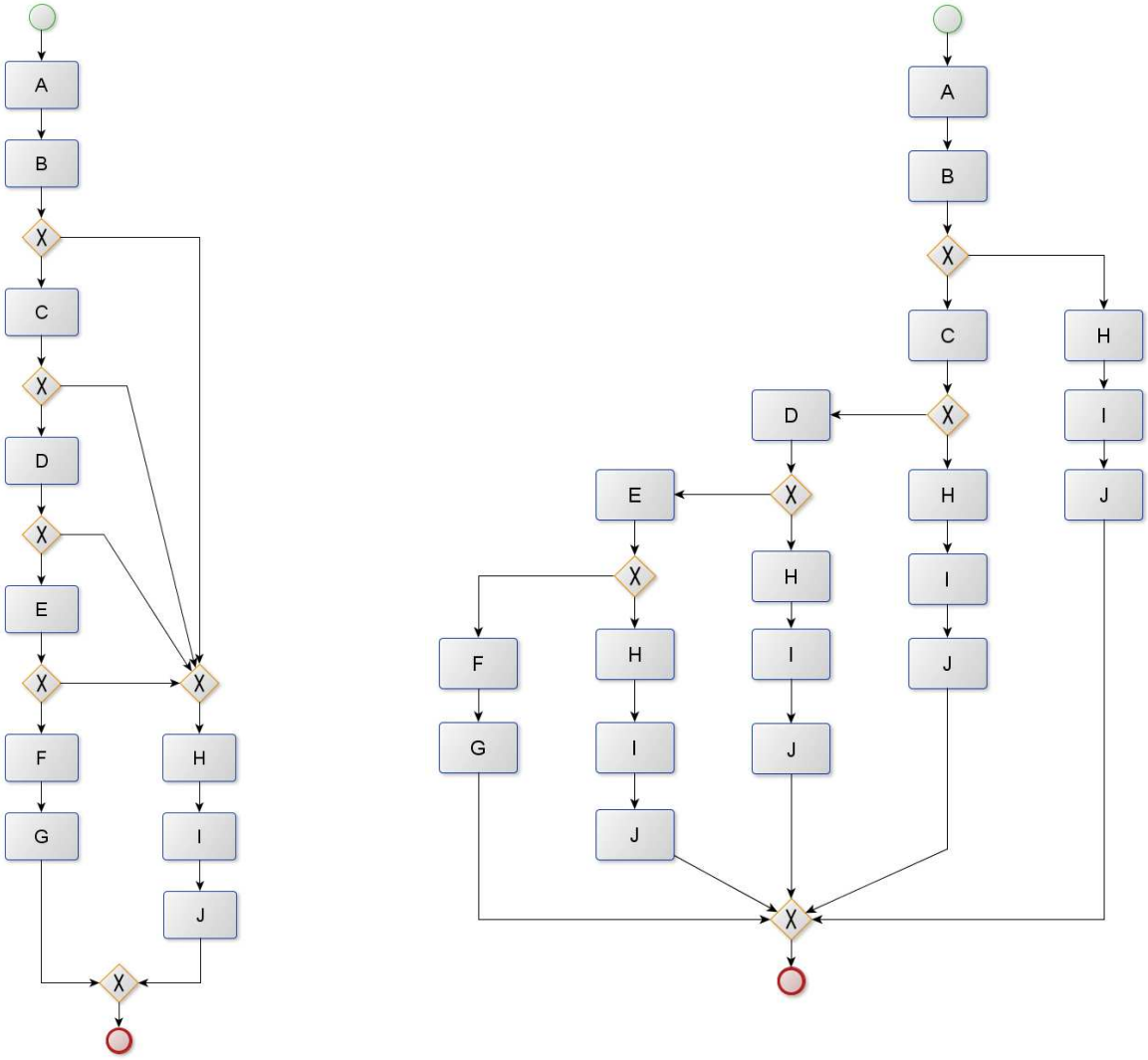
The CFC metric also has a very strong positive relation with all four size metrics. In case of #arcs and #GW it is 0,98, in case of #tasks it is 0,95 and finally in case of overall size it is 0,97. The statistical significance is below 0,007% for all of these correlations except #GW, which has a significance of 18%. This shows that the correlation between #GW and CFC is not statistically significant and the correlation between CFC and other size-related metrics is significant. Since it is an additive metric then it usually is higher in case of larger models. For that reasoning, CFC can be considered also as a size-metric and it has more meaning in case of models of relatively similar size. Comparing to other metrics, CFC has an average negative correlation with density (-0,5), slightly negative correlation with CC (-0,31) and no correlation with ACD (-0,07). For all of these correlations, the statistical significance is below 0,12% showing that the values are statistically significant. A high value of CFC usually corresponds with a low value of density, suggesting that there is a small contradiction between them as a complexity or understandability metric, where according to one the model is complex and according to the other the model is simple and the other way around.

CC has a small negative correlation with all of the size-related metrics. For each of #arcs, #GW, #tasks, size and CFC, the value of Pearson correlation is -0,31. The statistical significance of these values is below 0,12%, which shows that they are statistically significant. As high values of CC mean that the model should be less complex and easier to comprehend and low values of size should have the same meaning then these metrics support each other on a certain level. Comparing CC to ACD and density, then with those metrics, it has a positive correlation of 0,31 and 0,52 respectively. The statistical significance of these correlation values are 0,41% and 0%, which shows that they are significant. This is suggesting another contradiction between CC and density as this is a relatively strong correlation value and a lower value of density should relate to a more understandable model whereas for CC the same would apply for a higher value. With ACD the correlation is opposite to what was with size-related metrics, therefore creating another small contradiction.

For ACD, there is practically no correlation with size-related metrics. The correlation values are -0,06 for #arcs, 0 for #tasks, -0,07 for size and CFC. Only for #GW there is a small negative correlation of -0,2, which is caused by the fact that the calculation of ACD is based on the number of connectors. In all of these cases, the statistical significance is under 0,09%. However, there is a strong relation to density, which is shown by the Pearson correlation value of 0,71 with a significance value of 2,49%. This is caused by the fact that the two metrics have somewhat similar approaches in their calculation, as also stated above. For the relations between density and size-related metrics, the correlation values stay at around -0,5 with a significance value that is under 0,11% for all cases. In general, the density metric is in a contradiction with all other values, but ACD.

3.3 Examples of changes in the BP models

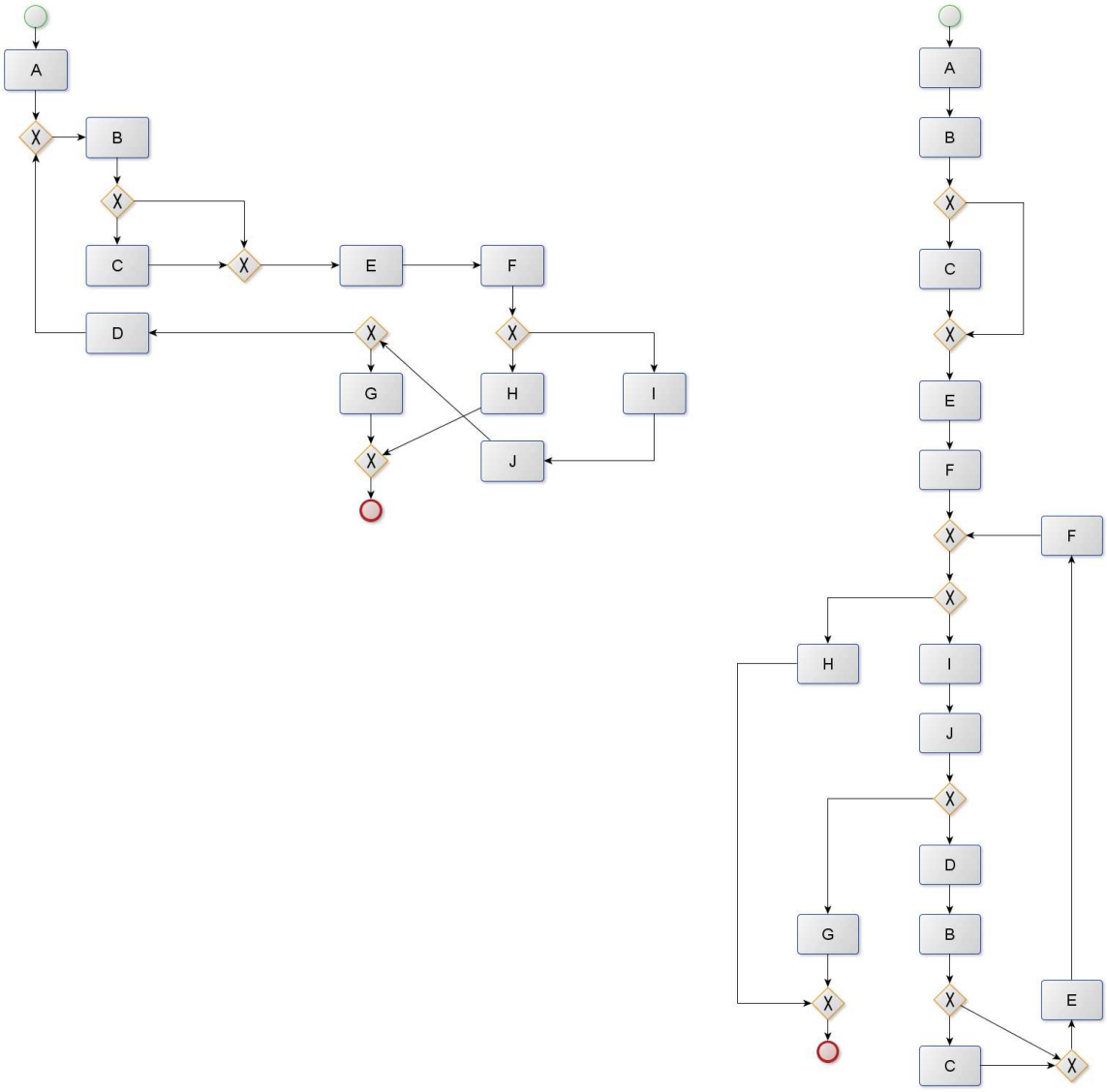
To illustrate some of the changes of the metrics, 5 examples of unstructured models and their structured counterparts will be presented in this section of the thesis. These examples will be selected from among the models that are used in the survey to empirically validate the results presented in this chapter. The 5 examples are ones for which there were more mistakes made in terms of their understandability in the survey. The survey and its results are described in further detail in Chapter 4.



(a) (b)

Figure 7: This example is Model 2 that was used in the questionnaire for the empirical study. Model (a) is unstructured and model (b) is structured.

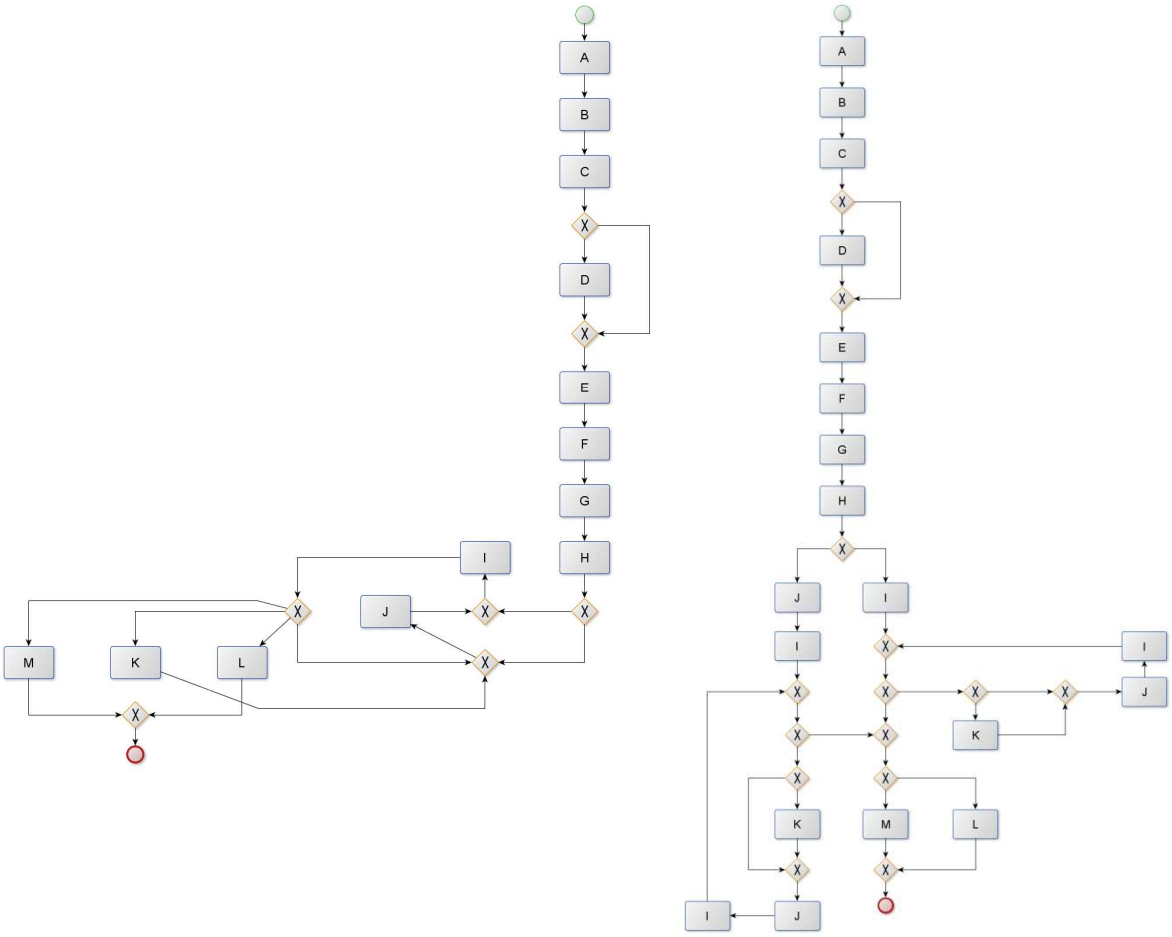
The structured version of Model 2 that can be seen in Figure 7(b) has 47,06% increase in size, 2,07% increase in CC, 14,55% increase in ACD and 17,14% decrease in density compared to its unstructured counterpart seen in Figure 7(a). For Model 2, we can see on the structured version of it, that there are 4 instances of the task sequence H, I, J, which explains the considerable increase in size. The final join gateway degree in the structured version has a higher than the largest join in the unstructured version and this explains the increase in ACD. CC does not change much and for density, it seems like the ratio of tasks to arcs is lower for the structured version.



(a) (b)

Figure 8: This example is Model 3 that was used in the questionnaire for the empirical study. Model (a) is unstructured and model (b) is structured.

The structured version of Model 3 that can be seen in Figure 8(b) has 35,29% increase in size, 122,90% increase in CC, 0% change in ACD and 11,11% decrease in density compared to its unstructured counterpart seen in Figure 8(a). Tasks B, C, E and F have been duplicated and two gateways have been added, therefore the size is larger for the structured version. CC is so much larger in this model, because the gateways are not so close together in the structured version. For CC, XOR-gateways have lower weight and the sum of values for all connections is calculated. If there are a lot of connections with gateways of large degree then the final value of CC will be smaller. ACD does not change and density changes for the same reason than in Model 2.



(a) (b)

Figure 9: This example is Model 4 that was used in the questionnaire for the empirical study. Model (a) is unstructured and model (b) is structured.

The structured version of Model 4 that can be seen in Figure 9(b) has 61,90% increase in size, 41,10% decrease in CC, 12,5% decrease in ACD and 44,79% decrease in density compared to its unstructured counterpart seen in Figure 9(a). In this model, size is increased, because tasks J, I and K have been duplicated several times and the number of gateways is also doubled. For the decrease in CC, this has happened for the exact opposite reason than what was provided for the change in CC for Model 3. ACD is smaller, because the degree of each gateway has been brought to a minimal of 3 in the structured version. Density decreases for the same reason as discussed before.

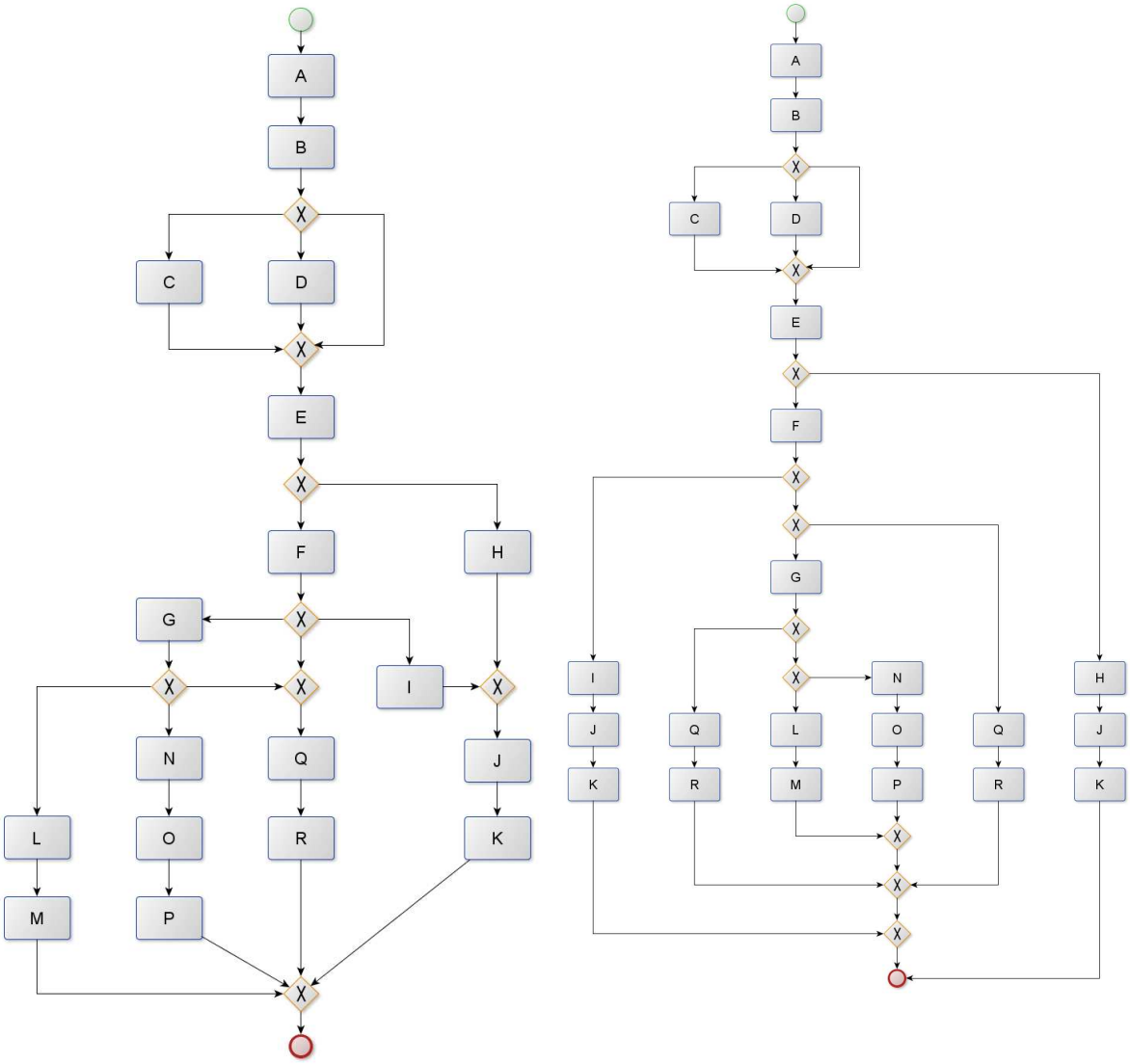
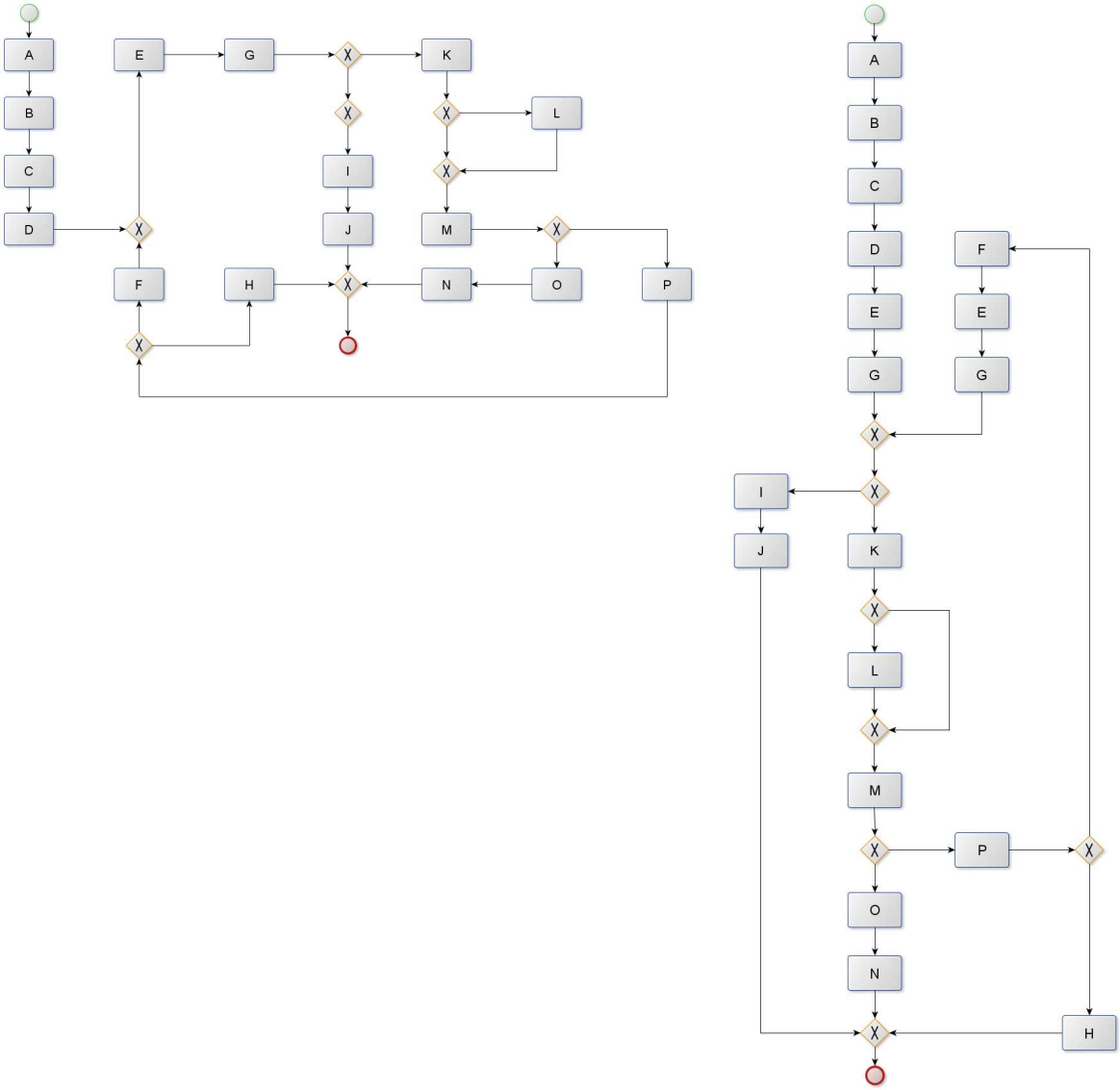


Figure 10: This example is Model 5 that was used in the questionnaire for the empirical study. Model (a) is unstructured and model (b) is structured.

The structured version of Model 5 that can be seen in Figure 10(b) has 22,22% increase in size, 2,62% increase in CC, 9,33% decrease in ACD and 26,00% decrease in density compared to its unstructured counterpart seen in Figure 10(a). For this model, size increases because tasks J, K, Q, R have been duplicated and two gateways have also been added. However, the task duplication is quite small comparing to the overall size of the model. CC is most probably a little larger, because in the unstructured version of the model, the last gateway has a large degree and every path through the model goes through the last gateway, which influences the final value of CC. Changes in ACD and density follow a similar pattern than for previous examples.

The structured version of Model 8 that can be seen in Figure 11(b) has 4% increase in size, 30,60% increase in CC, 4,76% increase in ACD and 6,06% increase in density compared to its unstructured counterpart seen in Figure 11(a). In this model, there is a very small change in size. Only two tasks, E and G, have been duplicated and one gateway has been removed. That gateway is the one with only one input and one output arc connected to it, rendering the gateway useless. Due to removing this gateway, the value of ACD is increased.



(a) (b)

Figure 11: This example is Model 8 that was used in the questionnaire for the empirical study. Model (a) is unstructured and model (b) is structured.

3.4 Conclusions

The results of the measurements can be approached from three standpoints. First are the size-related metrics. The five of them – #arcs, #GW, #tasks, size, CFC (this is added to the list of size-related metrics for the reasons stated in the previous chapter) – are very similar in their approach and the behavior of their values. From all of their values, we can interpret that the structured versions of the BP models are more complex than their unstructured counterparts. We can also make the assumption that a more complex model is more difficult to comprehend for an average person. This means that all of the metrics suggest a decrease in understandability of the restructured versions of the models. The average values of the metrics are 35 to 53 percent larger for structured BP models.

Secondly, we look at the CC metric, which behaves differently from other metrics. For CC, the average value decreased 11,44% in the structured versions of the models. Considering the definition of the metric, this suggests an increase in complexity and decrease in understandability. Considering that the maximum difference of this metric is 0,04711, then 11 percent may prove to be a big difference. CC metric also takes into account the existence of duplicated tasks, which could be an important factor considering the working principles of BPStruct.

Finally, there are ACD and density metrics, which are considered to be the best two metrics for evaluating BP model understandability in academic literature. ACD decreased 3,04% and density decreased 22,5% in the structured versions of the models. Both of these values indicate an improvement in understandability and therefore also a decrease in complexity. For density, the value of improvement could have a significant meaning.

However, for each of these models we do not have a proper scale for evaluating them except for CFC for which we can use the scale of the analogous McCabe's Cyclomatic Complexity metric that has been widely used in software engineering. For CFC, the scale suggests that indeed the restructured versions of the BP models are more complex. Since we do not have a scale to evaluate the other metrics, we need to validate the results by doing a controlled experiment where people assess both versions of the models by answering several questions about them.

In the next chapter of this thesis, a controlled experiment will be conducted among students of the University of Tartu who are taking a course on business process management. The chapter will introduce the design of the experiment, present the results of it and comparatively analyze the results with the results of the analytical study presented in this chapter. In the end of the next chapter, final conclusions will be made about the effectiveness of the BPStruct algorithm in whether the restructured models are easier to understand and less complex than their unstructured counterparts as is expected.

4. Controlled experiment

The aim of this chapter is to empirically validate the results presented in Chapter 3. In order to do this, a controlled experiment will be done among real people who will answer questions about unstructured models and their versions that were restructured with BPStruct. The chapter consists of four sections. The first one will introduce the controlled experiment and describe how it will be conducted. The second section will present the results of the experiment. The third section will analyze any factors that may affect the validity of the results presented in the second section. Finally, the fourth section will compare the results to the ones acquired in Chapter 3 and provide conclusions drawn from the experiment.

4.1 Experimental design

To gather the empirical data about the understandability of unstructured and structured BP models, a controlled experiment was conducted. A survey was made among a group of students that were split into two groups – a *control group* and an *experimental group*. The control group answered questions about unstructured versions of the models and the experimental group answered the same exact questions about equivalent structured versions of the models. This section will now further explain how the experiment was designed and conducted.

The groups of students answered the questions about models in the format of a survey. Two different versions of questionnaires were made – one version for the unstructured versions of the models and the second for the equivalent structured versions. 8 different models were selected to be used in the survey. The models were chosen to be drawn in BPMN notation. That was done because of the popularity of BPMN and because the students had all been taught BP modeling in that notation. The tasks of the models were labeled with simple abstract names like A, B and C. An example of a model can be seen in Figure 7. For each model, the students had to rate the perceived difficulty of the model on a 4-point scale and answer 4 yes/no questions about the model. In addition, every student answered 4 theoretical questions about modeling in BPMN in the beginning of the questionnaire.

The models were selected on the basis of their size and whether they had cycles or not. A cycle in a BP model means that some tasks can be visited twice. Based on these two factors, four categories of models were created:

1) Acyclic unstructured models whose equivalent structured models are at least 25% larger in size as the unstructured versions.
2) Cyclic unstructured models whose equivalent structured models are at least 25% larger in size as the unstructured versions.
3) Acyclic unstructured models whose equivalent structured models are of more or less the same size as the unstructured versions.
4) Cyclic unstructured models whose equivalent structured models are of more or less the same size as the unstructured versions.

For each of these four categories, four models were selected to be used in the questionnaires. Besides the factors used to create the categories, it was attempted to select the models with as different changes as possible in terms of CC, ACD and density. The values of the metric changes between unstructured and structured versions of the models can be seen in Table 5.

Table 5: Increase/decrease of size, CC, ACD and density in percentages.

	Category	Size	CC	ACD	Density
Model 1	1	+45,45%	+2,07%	+14,55%	-17,14%
Model 2	1	+47,06%	-7,63%	+8,00%	-21,43%
Model 3	2	+35,29%	+122,90%	0%	-11,11%
Model 4	2	+61,90%	-41,10%	-12,50%	-44,79%
Model 5	3	+22,22%	+2,62%	-9,33%	-26,00%
Model 6	3	+12,00%	+11,37%	+5%	+0%
Model 7	4	+15,79%	-9,97%	-4,55%	-18,75%
Model 8	4	+4,00%	+30,60%	+4,76%	+6,06%
Average		**+30,47%**	**+13,86%**	**+0,74%**	**-16,65%**

The questions asked about the model were designed in a similar fashion to the questions of BPMN-Selftest [23]. An example question from the questionnaire: “If J is executed in a

case, can F be executed in the same case?" In the BPMN-Selftest, the questions were presented one-by-one in a style where a model was displayed and under the model the current question was asked and the tasks in question were highlighted in the model. Since this experiment was conducted in a classroom and on paper, then exactly the same approach could not be used and all four questions about the model were displayed under it for reading and answering. Therefore the tasks in question could not be highlighted. In addition, there was a restriction about the size of the models. They had to be selected in a fashion where they would look clearly readable on a paper in A4 format while also leaving room for displaying 5 questions on the same page. Both questionnaires are included in the supplementary material of the thesis in the form of a companion CD.

The experiment was conducted among students of the University of Tartu, who were currently taking a course on business process management. The course is a part of an international software engineering master's degree program. This means that the students came from very different studying backgrounds. At that moment of time, they had learned the basics of modeling business processes in BPMN notation and had also done a homework where they had to model a process that was described in a case study. In total, 18 students answered the questionnaires, 9 of them were in the "control group" and 9 of them were in the "experimental group".

The students were given 20 minutes to answer the questionnaire, giving them about 30 seconds to answer each question, therefore not allowing them to lengthily concentrate on every question. They were reminded of the time limit and also a set of example questions was answered before doing the survey in front of the class in order to familiarize the students with the organization of the questionnaire. They were not given an overview of the exact specifics of the topic of unstructured vs. structured BP models and they were also not mentioned that in some of the models, tasks may be duplicated. In both groups, there was one student that could not finish the questionnaire on time, leaving some models unanswered. Since the missing data would have skewed the final results then only complete questionnaires were included in the final analysis. Therefore, a total of 72 answers to theoretical questions about modeling in BPMN, 144 evaluations of the perceived complexity of the model and 576 model-specific questions were given.

In the next section, the results of this survey will be presented and analyzed.

4.2 Results of the experiments

This section of the chapter will give an overview of the results of the controlled experiment that was introduced in the previous section. First, an analysis of average outcomes of the survey will be done. Then, a more detailed review will be given about separate questions for which there were most wrong answers given to by the students. It will be combined with examples of specific models. The full data of the question statistics is provided in the statistics spreadsheet in the supplementary material of the thesis.

4.2.1 Average results

The first part of the questionnaire was about theory of modeling in BPMN, where four questions were asked from each student. Every right answer to a question was worth one point. The average score for the control group was 1,5 and the average score for the experimental group was 2,25, resulting in a 50% better score for the students who were part of the latter group. The average results of the theoretical part of the questionnaire can be seen in Table 6.

Table 6: Results of the theoretical part of the questionnaire.

Group	Theory score on average
Control group (unstructured models)	1,5
Experimental group	2,25
Change in theory score	**50,00%**

Similarly to the theory part, an average score was also calculated for each model used in the survey. The students also had to evaluate their perceived complexity of every model on a 4-point scale. Therefore, for all of the models, the value of their perceived complexity was also calculated based on the answers given in the questionnaire. The average value of score over all models decreased by 3,91% for the restructured versions. At the same time, the average value of perceived complexity over all models increased by 14,68%. Therefore, even though that the students in the experimental group answered the theoretical questions better than the students in the control group, they evaluated the structured versions of the models as more complex and they also performed worse in

answering specific questions about the models. The average results for all models are summarized in Table 7.

The average decrease of score of 3,91% is quite small, however the scores were consistently lower for most models used in the survey. Out of 8 models, only 2 of them had an improvement in score and all of the other 6 showed a decline in score. The same happened with perceived complexity on a slightly larger scale. Firstly, the increase of 14,68% in complexity is a considerable growth. Secondly, out of the 8 models in total, for 7 of them, the perceived complexity was higher for the structured version. For the one model that a lower value, the average score was also lower. The two models that displayed an increase in score, the value of perceived complexity was rated as higher by the students.

Table 7: Change of score and perceived complexity of the models before and after restructuring.

	Change in score	Change in perceived complexity
Model 1	-6,45%	-13,33%
Model 2	+7,41%	+30,77%
Model 3	-8,70%	+4,76%
Model 4	+3,57%	+30,00%
Model 5	-3,45%	+16,67%
Model 6	-6,35%	+23,53%
Model 7	-9,38%	+10,00%
Model 8	-8,00%	+15,00%
Average	**-3,91%**	**+14,68%**

4.2.2 Analysis of questions with most incorrect answers

To approach this analysis, a criterion was needed for choosing the questions for which there were more incorrect answers than others. It was decided that if there was a question for which at least one of the two student groups that answered the questionnaire contained at least three people that answered the question wrong, then that question will be selected for analysis. In other terms, if there are less than 65% of correct answers for a question in either the unstructured or structured version of the questionnaire, then it will match the

criterion. This way the questions cover all of the following cases: 1) the score of the model is increased in the structured version, 2) the score of the model is decreased in the structured version and 3) the score of the model remains the same in both versions. In total, from 64 possible cases, there were 9 that matched this criterion. In this section, each one of those cases will be analyzed. These questions and the amount of correct answers to them in percentages are summarized in Table 8.

Table 8: Amount of right answers to questions selected for analysis in percentages.

	Percentage of correct answers (unstructured)	Percentage of correct answers (structured)	Percentage of correct answers (total)
Model 2 – Q1	100,00%	62,50%	81,25%
Model 2 – Q2	50,00%	100,00%	75,00%
Model 3 – Q1	37,50%	50,00%	43,75%
Model 3 – Q3	100,00%	62,50%	81,25%
Model 3 – Q4	50,00%	62,50%	56,25%
Model 4 – Q1	62,50%	75,00%	68,75%
Model 5 – Q2	62,50%	62,50%	62,50%
Model 8 – Q1	50,00%	62,50%	56,25%
Model 8 – Q3	87,50%	62,50%	75,00%

Model 2 – Q1. *If C is executed for a case, can J be executed for the same case? (Correct: yes)*

For this question, every student in the control group answered it correctly, but 3 students in the experimental group answered it incorrectly. If we look at the unstructured version of Model 2 on Figure 7(a), then we see that after locating J, we can simply backtrack from it and reach C very quickly. If we look at the structured version on Figure 7(b), there are four different instances of J there. It is possible that if the student first located the upper rightmost J on the model, she backtracked from it to see that both C and J come after an exclusive gateway and considering the time limit of answering the questions, she quickly answered the question according to what she noticed first.

Model 2 – Q2. *Is a case possible where F is executed after B? (Correct: yes)*

For this question, there were four students in the control group that answered this incorrectly and every student in the experimental group answered this correctly. There may be two reasons for why so many students made a mistake here. First is that the student followed the path, where the XOR-split after B leads directly to an XOR-join. From there, she saw that one cannot reach F from that point anymore and answered the question without looking further. The second possibility is that the four students misunderstood the question and thought that F has to be executed directly after B. However, in this case it would be strange that only students belonging to the control group made the same mistake.

Model 3 – Q1. *Is it possible to execute both H and J in the same case? (Correct: yes)*

For this question, there were 5 students in the control group and 4 students in the experimental group that answered it incorrectly, therefore making it the single most difficult question of the survey. If we look at the unstructured version on Figure 8(a), we can see that if we reach H in the execution flow, then the path only leads to the end node. However, if we reach J first, then we can go back to the beginning of the model through a loop and also reach H. If we look at the structured version on Figure 8(b), we can see that the flow is similar, and a loop can be taken to reach H after executing J. It is possible that once the students saw that both H and J came after an XOR-split, they did not notice the existence of the loop.

Model 3 – Q3. *Can you execute C after E is executed? (Correct: yes)*

For this question, every student in the control group answered it correctly, but 3 students in the experimental group answered it incorrectly. If we look at the structured version of the model, then we notice that beginning from the start node, there are C and E tasks that come in succession and cannot be executed the other way around. However, if we move further along, we notice that there are another C and E that are part of a loop. It is possible that the students who answered this question incorrectly did not notice the duplicated C and E and wrote their answer according to the first two matching tasks that they saw.

Model 3 – Q4. *Is it necessary to execute both D and B in the same case? (Correct: no)*

For this question, there were 4 students in the control group and 3 students in the experimental group that answered it incorrectly. In case of the control group, it is possible that the students approached this question from the bottom part of the model and located D first and saw that B can be reached right after D through an XOR-gateway. If this case is true then they did not notice the start event of the model. In case of the experimental group, it is evident from the model that B is a duplicated task. If the students first noticed the B in the upper part of the model then they probably answered this question correctly as the end event can be quickly reached after this B. However, if they first noticed the B in the lower part of the model it is probable that they answered the question incorrectly, because that B is executed directly after the D.

Model 4 – Q1. *If L is executed in a case, is it possible that K has also been executed in the same case? (Correct: yes)*

For this question, there were 3 students in the control group and 2 students in the experimental group that answered it incorrectly. If we look at the unstructured version of the model on Figure 9(a), we can see that K is part of a loop that can be used to reach L after K has been executed. However, the loop is well hidden among overlapping arcs and it is possible to miss it upon first look at the model. If we look at the structured version on Figure 9(b), we see that there are two loops that can be used to execute K before continuing to L. In this case it seems strange that the students have missed them and answered this question incorrectly.

Model 5 – Q2. *If J is executed in a case, can F be executed in the same case? (Correct: yes)*

For this question, there were 3 students in both groups that answered it incorrectly. Looking at the models on Figure 10(a) and 10(b) then one can assume that the students might have misunderstood the question and thought that F needs to be executed after J has been executed once.

Model 8 – Q1. *Can F and H be executed in the same case? (Correct: yes)*

For this question, there were 4 students in the control group and 3 students in the experimental group that answered it incorrectly. If we look at the unstructured version of the model on Figure 11(a), we can see that both F and H are part of a loop that has to be traversed twice in order to execute both tasks in question. It is likely that students did not think of the possibility of traversing the loop twice and take a different path in order to execute both tasks. For the structured version of the model, which can be seen in Figure 11(b), the loop does not have to be taken twice, but F is still a part of it. However, to reach the loop and H, one needs to go through an XOR-gateway, where one arc leads upwards and another downwards. This gateway might have been a confusing factor to some students.

Model 8 – Q3. *If M is executed for a case, can J be executed for the same case? (Correct: yes)*

For this question, there was 1 student in the control group and 3 students in the experimental group that answered it incorrectly. Looking at the structured version of the model, we can see that J is located upwards from M and in order to reach J, one has to go through the loop in upwards direction. For that reason, students may have missed seeing the loop. This was also the second to last question, which means there may have been some fatigue.

4.3 Threats to validity

The results of the controlled experiment strongly suggest that the structured versions of the BP models are more complex and therefore also more difficult to comprehend than their unstructured counterparts. However there are some factors that may affect the validity of the results of the controlled experiment. This section will analyze them and document the threats that may influence the validity of the experiment.

First of all, the controlled experiment was conducted in a very small scale. The number of people that answered the surveys was only 18 and only 16 of them were included in the

final results due to the fact that 2 students could not finish answering all of the questions about the BP models on time. The answers of these 2 people had to be removed from the final results, because the incomplete data could have skewed the results.

Another factor that may have influenced the final results of the experiment is that all of the subjects of the experiment were students. These students only had very little theoretical background in BP modeling and they had had very little practice with actual modeling. In [10] it is shown that people who have had more actual practice in creating BP models understand their specifics in greater detail and therefore understand more complex models easier. It can be argued that if the subjects of the controlled experiment had had more real-life experience in BP modeling, the results could have been different.

The next factor that could have affected the results is that the subjects were not aware that some of the models had duplicate tasks in them. They were all students that had been trained with simple models, which usually do not have duplicated tasks in them. Therefore, when answering a question about a model, if the subject recognized a task in question, she did not look further and based her conclusion to the question on the task that she noticed first. However, if the students had been told that there may be models with duplicated tasks, they could have been more thorough in studying the models. This might have affected the results in a different direction.

The last threat could also have been avoided if the tasks in question were highlighted. The survey was given to the students on paper and due to the limitations of an A4-sized paper, the tasks were not highlighted. If they had been, then the subjects would have had a clearer picture of the models and it would have been easier to find the right answer to the questions. Due to the same reasons, the size of the used models was also limited to a certain degree. A possibility to overcome these restrictions in the future could be using another medium to conduct the survey, e.g. an online survey. In this case, the models used can be larger in size and tasks in question can be highlighted. Larger models and highlighted tasks would allow using more complex models and a bigger variety of questions.

The final factors that could have affected the results of the controlled experiment are the drawing style of the models and the phrasing of the questions. In this case, the models were

drawn using the images of the original EPCs generated in ProM. Only change that was made was that the start node was moved to the topmost position and the end node was moved to the bottom of the diagram. Everything that was in the middle of those two nodes stayed relatively unaffected. However, looking at the models used in the survey in hindsight, it can be concluded that more changes in appearance could have been made that might make them seem more logical to a person examining them. Another point to consider is the phrasing of the questions used in the questionnaires. In the current survey, some of the questions did not have enough consistency in their phrasing. For example, a question phrased like “If A is executed for a case, can B be executed in the same case?” and a question phrased like “Can A and B be executed for the same case?” have a similar meaning. Given the limited amount of time the students had for answering the survey, these changes in phrases could have generated confusion. In order to eliminate that confusion, the questions need to have a similar style in phrasing.

The final section of this chapter will give a comparative overview of the results presented in Section 3.2 and the results presented in Section 4.2. It will also provide conclusions drawn from the results of the controlled experiment.

4.4 Conclusions drawn from the experiment

This section of the chapter will use Pearson correlation to find relations between score and perceived complexity values of the models in the survey and metrics calculated in Chapter 3 of the thesis. Based on the results presented in Section 4.2 and the result of the comparative analysis made in this section, final conclusions will be made about the controlled experiment.

To compare the data from chapters 3 and 4, Pearson correlation will be used similarly to how the metrics were compared in Section 3.2 of the thesis. The differences in score and perceived complexity of the unstructured and structured versions of the models used in the questionnaires will be compared to differences of size, CFC, CC, ACD and density metrics that were calculated for the same models to see how they are related between each other. The metrics of #arcs, #GW and #tasks will not be used due to their similarities to size and CFC. The values of the relations are summarized in Table 9.

Table 9: Relations between score and perceived complexity with size, CFC, CC, ACD, and density and their statistical significance values (in brackets).

Metric	Score	Perceived complexity
Size	0,63 (0,27%)	0,06 (0,4%)
CFC	0,45 (13,89%)	0,47 (32,08%)
CC	-0,4 (11,49%)	-0,4 (1,85%)
ACD	-0,1 (30,86%)	-0,6 (14,17%)
Density	-0,6 (21,91%)	-0,4 (1,41%)

Looking at the calculated values of the Pearson correlations, we can see that size has a somewhat strong positive correlation of 0,63 (with a significance of 0,27%) with the score of the model, meaning that a model that is larger in size scored higher in the empirical study. This indicates that size does not have as much weight as predicted in indicating the complexity or understandability of the model. The correlation value of 0,06 (with a significance of 0,4%) with the value of perceived complexity shows that these two values do not have a relation between them, indicating that subjects did not base their evaluations of the models upon the size of the model. Both of these correlations are statistically significant. With CFC, the correlation of 0,45 (with a significance of 13,89%) with model scores shows that also here the performance of the models was not related to this metric as a model with a higher CFC also got better results in score. Comparing CFC to perceived complexity gives a Pearson correlation value of 0,47 (with a significance of 32,08%). This shows that as opposed to size, CFC does indicate that students rated models that had bigger CFC value as more complex. However these values cannot be considered statistically significant. It is interesting to note that metrics that performed so similarly in the comparative analysis performed so differently in the empirical study.

In case of CC, the metric had a negative correlation of -0,4 with both the score of the models (with a significance of 11,4%) and perceived complexity of the models (with a significance of 1,85%), meaning that a lower value of CC resulted in a higher value of both score and perceived complexity in the empirical study. For perceived complexity, this shows a correspondence with the description of the metric, where it is stated that a decrease in CC implies an increase in error probability, meaning that a model with lower CC is generally more complex. However, with the score of the models, there is a

contradiction to this description. The Pearson correlation value of -0,4 indicates that a model with lower CC scored higher in the study than a model with a larger CC. According to the description of the metric, the correlation value should be positive. The correlation value with score is not and the value with perceived complexity is statistically significant.

In case of ACD, the metric had a negative correlation value of -0,1 (with a significance of 30,86%) with the scores of the models, indicating that ACD did not have much influence to the results of the survey. The correlation value of -0,6 (with a significance of 14,16%) with the perceived complexity of the models suggests that subjects of the survey rated models that had a lower ACD as more complex, which once again creates a contradiction between the results of the survey and the nature of the metric. However, neither of these correlation values can be considered statistically significant). For density, the Pearson correlation value of -0,6 (with a significance of 21,91%) with score of the models suggests that for a model with lower density, the values of score were higher, therefore confirming that the metric does indeed indicate understandability of the models. However, for perceived complexity, the small negative correlation of -0,4 (with a significance of 1,41%%) with density also shows that subjects rated models with lower density as more complex. This creates a contradiction between density and perceived complexity. The correlation value with score is not and the value with perceived complexity is statistically significant.

Summarizing the results of the comparison between the analytical metrics and the empirical study, it can be seen that there are a lot of contradictions between what was predicted to happen and what actually happened. Only in three cases was it seen that the predicted behavior of the metric matched the actual behavior. Density metric confirmed that there is a relation between it and understandability of models however the value cannot be considered statistically significant. CFC and CC metric showed a small relation between them and the perceived complexity of the model with CC also being significant. All other combinations had either a contradictive relation or did not show any relation at all. For example, the metric of size does not seem to have any impact on understandability. However looking at the average results of the survey, the opposite could have been predicted as every structured model used in the survey was larger in size than its unstructured counterpart.

From this, it can be concluded that there seem to be many other factors that influence the complexity and understandability of the models. Unlike the metrics used, these factors may be immeasurable. As stated in Section 4.3, factors that may influence the understandability of the models could be model appearance, subjects' understanding of the questions or the fact that in some models there existed duplicated tasks about which the subjects were not aware of. Another aspect that probably influenced these results and did not allow presenting a clearer picture was that there were very few models used in the survey, the size of those models was restricted, and very few subjects took part of the survey.

From the aggregated average results, we can make the following three observations. The subjects in the experimental group answered theoretical questions about modeling in BPMN considerably better, suggesting that they also understand the nuances of BP modeling better. They also rated the perceived complexity of the structured models as more complex than the control group did for the unstructured versions of the models. Finally they answered specific questions about the structured BP models worse than the control group did for the unstructured counterparts of those models. From this we can conclude that the research hypotheses H1 and H2 cannot be confirmed. Models restructured with BPStruct are currently not less complex or easier to comprehend than their unstructured counterparts.

To further analyze these results, alternative possibilities have to be considered as to why the models restructured with BPStruct seem to be more complex and less understandable. In [24], it is argued that there may be good reasons to intentionally create BP models that are unstructured. The authors are building a catalogue of different patterns of unstructuredness and are categorizing them as being good or bad reasons of having unstructured elements in BP models. Looking at the structured BP models that are obtained by transformation from unstructured versions, we see a lot of duplicated tasks in them, which adds complexity to the models and generates confusion when reading the modes. This may well be a reason for BP modelers with lots of real-life experience in BP modeling to knowingly create some models as unstructured, since the alternative structured versions would have a considerable amount of concurrency in them.

The final chapter of the thesis will give a short overview of the results presented in the thesis and review the conclusions made. It will also discuss how this topic can be further addressed in the future.

5. Conclusions

This thesis focused its research on studying the problem of the complexity and understandability differences of unstructured and structured BP models and tried to find answers to the research questions of whether restructured versions of originally unstructured models are simpler in terms of complexity and whether they are easier to comprehend. The reviewed literature on BP modeling suggested strongly that it is desirable that models follow some structural rules and modelers use it as a guideline to design their models as structured as possible. This gave reason to assume that the answer to both research questions is “yes”. Therefore, the following hypotheses were formulated:

- H1. BP models restructured with BPStruct are less complex than equivalent unstructured ones;
- H2. BP models restructured with BPStruct are easier to comprehend than equivalent unstructured ones.

To approach these hypotheses, firstly, the concept of unstructuredness was introduced and the basic working principles of a restructuring program called BPStruct were explained. Then, an overview was given about existing research on measuring BP models in terms of complexity and understandability and a set of metrics was selected to be used in this study. These metrics were then used to perform a comparative complexity study on a dataset of real-life BP models and their equivalent structured counterparts. In an attempt to empirically validate the results of the study, a controlled experiment was conducted using students taking a course on BP management to evaluate both unstructured and structured versions of the models.

The metrics used in the comparative complexity study were number of arcs, number of gateways, number of tasks, size of the model, control-flow complexity, cross-connectivity, average connector degree and density. The size-related metrics, CFC (which could also be considered to be a size-related metric) and CC are primarily complexity metrics and average connector degree and density can be considered as understandability metrics as academic research has shown that they are two most powerful metrics in evaluating BP model understandability. On average, the structured versions grew considerably in size,

they had a larger CFC and a smaller CC, showing that the structured versions gained in complexity compared to the unstructured ones. On the other hand, they also decreased in both ACD and density, which means that the metrics show that the structured versions should be more understandable.

In the controlled experiment to empirically validate the results there were 8 subjects in the control group who answered questions about unstructured versions of the models and 8 subjects in the experimental group who answered the same questions about the structured versions of the models. The result shows that even though the experimental group answered theoretical questions about modeling in BPMN better, they answered specific questions about BP models worse than the control group. They also evaluated the models as more complex than the control group did. These results show that the structured versions are both more complex and also more difficult to understand than the unstructured ones. In terms of the metrics from the comparative study, then the only strong and statistically significant relation was between size and model scores, which quite interestingly shows that the mistakes in answering the questions about models were not dependant on the size of the model.

Taking into consideration both the results from the comparative complexity study and the controlled experiment then we can deny both hypotheses H1 and H2 that were formulated. First of all, for H1, it can be said that the BP models restructured with BPStruct are not less complex than their unstructured counterparts. Secondly, for H2, it can be said that the BP models restructured with BPStruct are not easier to comprehend than their unstructured counterparts.

However, with these results, the threats to validity that were documented in Section 4.3 have to be taken into consideration. The amount of models used in the survey was small and also the number of subjects that took part of the experiment was also very small. This means that strong claims cannot be made about the results. The study was also conducted using questionnaires on paper, which meant that the A4 paper format presented its own limitations in terms of model selection. However, the results of the study suggest that when process modelers draw unstructured process models, an assumption can be made that they do it for the right reasons, in the sense that the corresponding structured BP models are less readable and more error-prone. That is because the structured versions contain event

duplications which generate confusion when reading the models and drawing BP models that are larger in size could lead to the modeler making more errors in designing the model. This means that drawing the models as unstructured is the only alternative.

To expand on the results of this study in the future, a more comprehensive controlled experiment can be conducted. As stated, the experiment done within this thesis contained too few subjects and too few models were analyzed during it. The experiment was also limited due to it being conducted on paper. In order to improve on it, the study has to be done through a different medium, e.g. an online survey. With an online survey, more people can be involved, more models can be used, tasks that are used in the questions can be highlighted on the displayed BP model, etc. The survey could also benefit from identifying the good and bad patterns of unstructuredness in BP models and using it as a reference point in evaluating the results between unstructured and structured versions of the models.

Keerukuse ja arusaadavuse võrdlus struktureerimata ja struktureeritud äriprotsessimudelite vahel

Magistritöö
Raul Mäesalu

Resümee

Käesoleva magistritöö peamine eesmärk on välja selgitada, kas struktureerimata kujul olevate äriprotsessimudelite transformeerimine struktureeritud kujule muudab nad vähem keerukamaks ning lihtsamini arusaadavamaks. Püstitatud hüpoteeside järgi on struktureeritud kujul mudelid keerukuselt lihtsamad ning kergemini arusaadavad.

Töös kasutatakse varasemas uurimistöös valminud avatud lähtekoodiga programmi BPStruct, mille abil transformeeritakse hulk äriprotsessimudeleid struktureeritud kujule. Kasutatakse reaalsest elust pärit mudelitest koosnevat IBM andmestikku. Nimetatud mudelid mõõdetakse akadeemilises kirjandusest kirjeldatud meetrikate põhjal ning viiakse läbi võrdlev uurimus.

Võrdleva uurimuse käigus saadud tulemusi kontrollitakse eksperimendi abil, mille käigus Tartu Ülikooli tudengid, kes õpivad ainet nimega Äriprotsesside juhtimine, jaotatakse kahte gruppi – kontrollgrupp ja eksperimentaalne grupp. Kontrollgrupi tudengid vastavad struktureerimata kujul olevate mudelite kohta spetsiifilisi küsimusi. Eksperimentaalse grupi tudengid vastavad samade mudelite struktureeritud kujul olevate variantide kohta samadele küsimustele.

Nimetatud kahe uurimuse tulemuste vahel viiakse läbi võrdlev analüüs ning selle põhjal tehakse järeldused selle kohta, kas struktureeritud kujul olevad mudelid on tõesti lihtsamad ning kas neist on kergem aru saada.

References

[1] Definition of “business process” on SearchCIO.TechTarget.com. http://searchcio.techtarget.com/definition/business-process (17.05.2011)

[2] Volker Gruhn, Ralf Laue: What business process modelers can learn from programmers. Sci. Comput. Program. 65(1): 4-13 (2007)

[3] Jan Mendling, Hajo A. Reijers, Wil M. P. van der Aalst: Seven process modeling guidelines (7PMG). Information & Software Technology 52(2): 127-136 (2010)

[4] Artem Polyvyanyy, Luciano García-Bañuelos, Marlon Dumas: BPStruct on Google Project Hosting. http://code.google.com/p/bpstruct/ (17.05.2011)

[5] Artem Polyvyanyy, Luciano García-Bañuelos, Marlon Dumas: Structuring Acyclic Process Models. BPM 2010: 276-293

[6] Artem Polyvyanyy, Luciano García-Bañuelos, Marlon Dumas: Unraveling Unstructured Process Models. In: Proceedings of the 2nd International Workshop on the BPMN (2010)

[7] Artem Polyvyanyy, Luciano García-Bañuelos, Marlon Dumas: Structuring Acyclic Process Models presentation slides. http://www.slideshare.net/ArtemPolyvyanyy/structuring-acyclic-process-models (17.05.2011)

[8] Boudewijn F. van Dongen, Jan Mendling, Wil M. P. van der Aalst: Structural Patterns for Soundness of Business Process Models. EDOC 2006: 116-128

[9] Irene Vanderfeesten, Jorge Cardoso, Jan Mendling, Hajo A. Reijers, Wil van der Aalst: Quality Metrics for Business Process Models. In: Fischer, L. (ed.) BPM and Workflow Handbook 2007 (May 2007), pp. 179-190.

[10] Jan Mendling, Hajo A. Reijers, Jorge Cardoso: What Makes Process Models Understandable? BPM 2007: 48-63

[11] Jan Mendling, Mark Strembeck: Influence Factors of Understanding Business Process Models. BIS 2008: 142-153

[12] Jorge Cardoso, Jan Mendling, Gustaf Neumann, Hajo A. Reijers: A Discourse on Complexity of Process Models. Business Process Management Workshops 2006: 117-128

[13] Jorge Cardoso: How to Measure the Control-flow Complexity of Web Processes and Workflows. In: Fischer, L., ed., Workflow Handbook 2005, pp. 199-212

[14] Irene T. P. Vanderfeesten, Hajo A. Reijers, Jan Mendling, Wil M. P. van der Aalst, Jorge Cardoso: On a Quest for Good Process Models: The Cross-Connectivity Metric. CAiSE 2008: 480-494
[15] Jan Mendling: Testing Density as a Complexity Metric for EPC's. Technical Report JM- 2006, 11-15. 2006
[16] Dirk Fahland, Cédric Favre, Barbara Jobstmann, Jana Koehler, Niels Lohmann, Hagen Völzer, Karsten Wolf: Instantaneous Soundness Checking of Industrial Business Process Models presentation slides.
http://www.informatik.uni-rostock.de/~nl/wiki/publications/fahlandfjklvw_2009_bpm (17.05.2011)
[17] Dirk Fahland, Cédric Favre, Barbara Jobstmann, Jana Koehler, Niels Lohmann, Hagen Völzer, Karsten Wolf: Instantaneous Soundness Checking of Industrial Business Process Models. BPM 2009: 278-293
[18] Business Process Model and Notation (BPMN) Version 2.0.
http://www.omg.org/spec/BPMN/2.0/ (17.05.2011)
[19] Jan Mendling, Markus Nüttgens: EPC markup language (EPML): an XML-based interchange format for event-driven process chains (EPC). Inf. Syst. E-Business Management 4(3): 245-263 (2006)
[20] ProM webpage.
http://www.processmining.org/prom/start (17.05.2011)
[21] Grep webpage.
http://www.gnu.org/software/grep/ (17.05.2011)
[22] Frederick J. Gravetter, Larry B. Wallnau: Statistics for Behavioural Sciences. Wadsworth Publishing; 008 edition (December 10, 2008)
[23] Jan Mendling, Johannes Wolf, Hajo A. Reijers, Matthias Schrepfer: BPMN-Selftest - The Survey on Understanding of BPMN Process Models.
http://www.bpmn-selftest.org/ (17.05.2011)
[24] Volker Gruhn, Ralf Laue: Good and Bad Excuses for Unstructured Business Process Models. In: Proceedings of 12th European Conference on Pattern Languages of Programs (EuroPLoP 2007).

Supplementary Material

The companion CD attached to this thesis contains:

- IBM dataset. Located in the folder named “Dataset” on the CD. The folder contains all original models of the dataset and restructured models of the dataset.
- Questionnaires. Located in the folder named “Questionnaires” on the CD. Contains both versions of the questionnaire – unstructured and structured version. Also contains the answers to the questions presented in the questionnaires.
- Spreadsheet of statistics and measurements. Located in the folder named “Statistics”. Contains results of the comparative complexity study, results of the controlled experiment and numerical analysis of the questions with most incorrect answers.