The Journal of Systems & Software 144 (2018) 143–164

Contents lists available at ScienceDirect

The Journal of Systems & Software

journal homepage: www.elsevier.com/locate/jss

Artifact-based vs. human-perceived understandability and modifiability of refactored business processes: An experiment

Danilo Caivano[a,d], María Fernández-Ropero[b], Ricardo Pérez-Castillo[c,*], Mario Piattini[c], Michele Scalera[a]

[a] Department of Informatics, University of Bari, Via E. Orabona, 4, 70126 Bari, Italy
[b] Indra Software Labs, Ronda de Toledo, s/n, 13005 Ciudad Real, Spain
[c] Information Technology & Systems Institute, University of Castilla-La Mancha, Paseo de la Universidad 4, 13071, Ciudad Real, Spain
[d] SER&Practices s.r.l. Spin Off Company of the University of Bari, Via E. Orabona, 4, 70126 Bari, Italy

ARTICLE INFO

Keywords:
Business process
Refactoring
Understandability
Modifiability
Empirical study
Controlled experiment

ABSTRACT

Business processes modeling has proven to be effective and reverse engineering techniques with which to recover business process models when they are missing or outmoded have therefore emerged. Regrettably, these techniques often lead to models with quality flaws and consequently to models with low levels of understandability and modifiability. Refactoring has been widely used to deal with such flaws, altering the internal structure of models while preserving their semantics. There are several studies concerning how understandability and modifiability are affected by refactoring in terms of several artifact-based measures. However, there is little evidence regarding how refactoring affects quality in terms of human-perceived measures. The goals of this paper are, therefore: to collect further empirical evidence about the influence of refactoring on understandability and modifiability of business process models and to investigate the correlation between artifact-based understandability and modifiability and human-perceived ones. The obtained results are not trivial and show that business process obtained by means of reverse engineering has recurrent quality flaws, and the understandability and modifiability of business process models cannot be assessed by using artifact-based measures only. Human-perceived measures need to be taken in to consideration in order to have a more accurate evaluation.

1. Introduction

Business process modeling allows us to understand the business activities that an organization carries out. Business process models provide a representation of an enterprise and depict the system functionality through the description of all its components and the interactions between them, in addition to describing the resources and goals involved (Weske, 2007). These models follow standard notations such as BPMN (*Business Process Modeling and Notation*) (OMG, 2011) in order to be understandable by stakeholders.

Business process modeling provides several benefits for both, enterprise management and software development. Despite all these benefits, some organizations have never carried out their own business process modeling, or it may be that their business process models are outdated and misaligned with regard to actual daily operation. It is for this reason that reverse engineering techniques have emerged in an attempt to retrieve business process models from existing source code or event logs (Di Francescomarino et al., 2009; R. Pérez-Castillo et al., 2011; Zou and Hung, 2006; Bianchi et al., 2000). Although reverse engineering is perceived as less error-prone and time-consuming than manual modeling, it often leads to some quality flaws that emerge as a consequence of the low abstraction level of the reconstructed models: redundancies (e.g., the same element is retrieved twice from two different elements in code); irrelevancy (e.g., an element that is not related to a business activity is abstracted from code); inconsistency (e.g., a business process element is retrieved in an isolated form and without some of the required relationships); and so forth.

Cutting-edge techniques like merging, mining, refactoring, re-use, among others, have been designed in recent years in an effort to deal with these quality problems (Dijkman et al., 2012). Refactoring in particular has been used by several authors in literature in the quest to improve the degree of quality in business process models. Refactoring techniques consist of changing the internal structure of business process models without altering or modifying their external behavior, and a

[*] Corresponding author.
E-mail addresses: danilo.caivano@uniba.it (D. Caivano), msfernandezr@indra.es (M. Fernández-Ropero), ricardo.pdelcastillo@uclm.es (R. Pérez-Castillo), mario.piattini@uclm.es (M. Piattini), michele.scalera@uniba.it (M. Scalera).

https://doi.org/10.1016/j.jss.2018.06.026
Received 31 October 2017; Received in revised form 1 June 2018; Accepted 7 June 2018
Available online 15 June 2018

refactoring operator therefore replaces some fragments with equivalent ones. Several refactoring operators with which to recognize refactoring opportunities and then apply different refactoring transformations have been proposed in literature (Dijkman et al., 2012; Weber et al., 2011; Dijkman et al., 2011; La Rosa et al., 2011; Leopold et al., 2010; Gambini et al., 2011; M. Fernández-Ropero et al., 2013). In addition, there is a proposal (La Rosa et al., 2011) especially designed to refactor business process models retrieved by means of reverse engineering.

The quality flaws mentioned have to be addressed in business process models, since these faults affect understandability and modifiability. These quality characteristics have proven to be two of the most challenging characteristics to consider in business processes (L. Sánchez-González et al., 2010; Reijers and Mendling, 2011). According to the international standard for the quality of software products ISO/IEC 25,010 (ISO/IEC, 2011), understandability represents the degree to which users recognize whether the product is appropriate for their needs. Modifiability, on the other hand, is the degree to which a business process model is effectively and efficiently modified without introducing defects or degrading performance. Business process models with adequate levels of quality make it possible to take advantage of the aforementioned benefits.

Since understandability and modifiability can be considered as extrinsic quality characteristics, they are difficult to evaluate without human intervention. Some studies such as (L. Sánchez-González et al., 2010; L. Sánchez-González et al., 2010) have analyzed the relationships between (i) certain intrinsic measures and indicators (e.g., size, connectivity, separability, density or depth) that can be directly quantified from business process models, and (ii) the gain in understandability and modifiability obtained after refactoring. For example, according to this kind of studies, it can be stated that a smaller business process model is theoretically more understandable. These “artifacts-based” studies are a means of assessing the understandability and modifiability of business process models before and after refactoring, without the time-consuming intervention of humans. Despite the fact that the authors of all the aforementioned works conducted empirical evaluations with students/practitioners to assess how metrics such as size affect the human beings‘ perceived understandability of process models, an in-depth assessment is necessary to demonstrate that these measures are really related to the perceived understandability and modifiability. Thus, the research questions that this paper addresses are:

1) Can refactoring improve the modifiability and understandability of business process models?
2) Is there a correlation between artifact-based measures and human-perceived ones?

In order to answer to them the paper reports the results of an experiment, involving 65 students, aimed at investigating the relationship between the artifact-based and human-perceived understandability and modifiability of business process models. The contribution of this paper is twofold:

1) The collection of empirical evidence that refactored business process models are better understood and easier to modify, while on the other, the time spent performing the understandability and modifiability tasks decreases with refactored models. Effectiveness and efficiency are therefore improved by using refactoring.
2) The evidence of the existence of a relationship between artifact-based measures related to the understandability and modifiability assessment (such as size, connectivity, separability, density and depth) and human-perceived ones. The correlation between artifact-based and human-perceived understandability and modifiability is negative with regard to size and depth, as previous works propose (e.g., the greater the size, the worse the understandability). Connectivity and density, meanwhile, have a positive correlation, while separability has a negative correlation, thus contradicting previous assumptions (L. Sánchez-González et al., 2010; Mendling et al., 2007). Nonetheless, the degree of correlation is weak and it is, therefore, impossible to draw strong conclusions.

The remainder of this paper is organized as follows: Section 1 presents the background by summarizing some related works. The subsequent sections present an in-depth empirical study carried out by means of a controlled experiment with the objective of obtaining some insights into the effect of refactoring on business process models, especially those retrieved by using reverse engineering. The experiment is based on the formal protocol with which to conduct and report empirical research in software engineering proposed by Jedlitschka et al. (2008). In accordance with this protocol, Section 3 shows how the experiment was planned and provides all the information needed to replicate it. The execution of the experiment is described in Section 4, while Section 5 sets out the entire data analysis, the discussion of which is provided in Section 6. Finally, Section 7 presents the conclusions drawn, along with future steps to be taken.

2. Background

Business process modeling and management have proven to be of great benefit enterprise modeling, as well for and software development. Several reverse engineering techniques with which to support business process recovery (Normantas and Vasilecas, 2013) have therefore emerged. However, these techniques imply the abstraction of information, and semantics are very often lost (Canfora et al., 2011); as a consequence, retrieved business process models frequently have quality faults such as missing or non-relevant elements, fine-grained elements, uncertainties and ambiguities (M. Fernández-Ropero et al., 2013). Fixing quality faults and improving business process models are topics that have been discussed by several authors in the last few years. Dijkman et al. (2012) provide several techniques such as merging, mining, refactoring or re-use, with refactoring being the technique most widely used by authors in literature. For instance, Weber et al. (2011) collect a catalogue of process model *smells* for the identification of refactoring opportunities. Dijkman et al. (2011) contribute by showing a technique that is based on metrics with which to detect refactoring opportunities. Similarly, La Rosa et al. (2011) identify patterns for the reduction of model complexity using means that include compacting, compositing, and merging. Dumas et al. (2011) and Ekanayake et al. (2012), meanwhile, focus on the detection of duplicate fragments (also called *clones*). Other authors, like Leopold et al. (2012), focus on the refactoring of activity labels in a business process model, following a verb-object style. Pittke et al. (2013) also focus on labels through the definition of a mechanism that can be used to identify synonym and homonym labels in model repositories. In an effort to retain relevant information, other approaches such as Smirnov et al. (2012,2011), Polyvyanyy et al. (2010), Smirnov (2012) pay attention to the identification of coarse-grained activities by means of business process abstraction, omitting anything that is insignificant. Conforti et al. (2014) focus on both discovering sub-processes in BPMN models and interrupting and non-interrupting boundary events and activity markers.

All of the above approaches are intended to be used with business process models discovered by employing mining process, e.g., using event logs as also occur in van der Aalst (2012) or by hand (Indulska et al., 2009). Other authors, such as M. Fernández-Ropero et al. (2013), Pérez-Castillo et al. (2014), and Caivano (2005), Caivano et al. (2001), attempt to identify and address quality challenges in business process models retrieved by means of reverse engineering. They define a technique and framework, IBUPROFEN, with which to refactor business process models specifically retrieved by using reverse engineering, in line with the BPMN notation. Their proposal allows different refactoring operators to be applied, considering their behavior: maximization of relevant elements, fine-grained

Table 1
Experiment overview.

Goals	1) Analyze refactored and non-refactored business process models with the purpose of evaluating them with respect to their understandability and modifiability, from the point of view of the researchers, in the context of a university course in software engineering with bachelor students. 2) Analyze artifact-based and human-perceived measures collected with the purpose of evaluating their relationships with the understandability and modifiability, from the point of view of the researchers, in the context of a university course in software engineering with bachelor students.
Experimental subjects	65 Computer Science students from the University of Bari.
Experimental units	Five business process models retrieved from two real-life information systems using reverse engineering together with the refactored versions.
Tasks	Artifact and human-perceived questions to assess the understandability and modifiability of each business process model.
Hypotheses	There is no significant difference in understandability and modifiability between refactored and non-refactored business process models. There is no correlation between the artifact-based and human- perceived measures for Understandability and Modifiability assessment
Variables	Ratio of correct answers to understandability and modifiability questions and time spent; Likert scale reporting perceived effects.
Design	Within-subjects design based on two groups of students. Each student understands and modifies artifacts and then fills in five questionnaires appertaining to business process models to which refactoring has and has not been applied.
Procedure	Background lecture, pre-test, experiment and post-test
Analysis	Mann-Whitney and Spearman test

granularity reduction and completeness.

The quality gain of a refactored business process model is difficult to measure, since it also depends on the people in charge of using, managing or evaluating these business process models; this is subjective and varies according to the particular individual/s involved. The papers previously referred have conducted some empirical validations of their proposals mainly through quantitative analysis of certain measures related to quality features. For example, Pérez-Castillo et al. (2013) present an artifact-based empirical study concerning the effect of refactoring on business process model understandability, in which a set of measures is used to quantify this effect but human perception is not considered.

Quality measurement in business processes has been addressed by several authors. Rolon et al. (2009) presented a set of measures with which to evaluate the structural complexity of a business process model in accordance with the BPMN notation at a conceptual level. The number of events, the number of gateways, and the number of association flows, among other aspects, were considered by these authors as measures that could be used to evaluate how understandable a business process model is. Similarly, L. Sánchez-González et al. (2010) presented a systematic review in which many measures for business processes were defined and applied to models. They also analyzed process model quality from the perspective of understandability and modifiability and determined threshold values in order to distinguish between different levels of process model quality (L. Sánchez-González et al., 2010; Sánchez-González et al., 2013). Zugal et al. (2012) also study the assessment of model understandability, but they focus on modularity and its impact on models. These authors start with the assumption that the hierarchy is not beneficial to the understandability of the model. However, the empirical evaluation of this work is still lacking. Factors that have an influence on the understandability of a business process model have been addressed by other authors, such as Mendling et al. (2007), Mendling and Strembeck (2008). Both groups of authors coincide in that they consider measures such as size, separability, diameter, etc. in order to assess the understandability of a business process model. Moreover, all of the above studies analyze the relationships that some measures and indicators have with regard to the modifiability and understandability. For example, a smaller business process model is theoretically more understandable. Appendix I shows each measure and its association with the characteristics of understandability and modifiability, as reported by the authors mentioned above.

These artifact-based studies have, therefore, been a means of assessing the quality of business process models before and after refactoring, without the time-consuming intervention of human beings. Unfortunately, the low presence of human opinion in those studies makes it impossible to demonstrate that these measures are really related to the understandability and modifiability perceived.

A validation with human beings is therefore required in order to validate refactoring techniques and verify whether there is a relationship between the hypothesized understandability/modifiability (using the measures presented in Appendix I - Table 16) and the understandability/modifiability perceived by humans.

Human-perceived validations are normally carried out during research by using experiments with students owing to the difficulty involved in doing so with a large number of professionals. Authors such as Nugroho, (2009) analyzed the understandability of UML diagrams with different levels of detail in the development phase. To address his hypothesis, *Nugroho* uses students from the Eindhoven Technology University. Similarly, Abrahão et al. (2011) present an experiment with students as a means to test their proposal. Another case of the use of students in software experiments is the work of Ricca et al. (2007). These authors assess the effectiveness of UML stereotypes for Web design to provide support for comprehension tasks.

3. Planning the experiment

The purpose of this section is to provide a description of the plan or protocol that was then used to perform the experiment and to analyze the results. For this purpose, authors followed guidelines for experimentation in software engineering proposed by Wohlin et al. (2012) and Juristo and Moreno (2013). Additionally, authors followed the recommendations of Jedlitschka et al. (2008) for reporting the experiment, also keeping in mind indications outlined in Ferreira et al. (2018). A replication package for this research (Baldassarre et al., 2014; Carver et al., 2013) is provided via the relevant URL (M. Fernández-Ropero et al., 2013).

This section first presents the goals, the experimental units, the experimental material and experiment tasks, the hypotheses of the experiment, the variables, the experiment design, the procedure and the analysis procedure. The aim is to provide all the details needed to make the experiment replicable. Table 1 represents an overview of the aforementioned items, which are then addressed in the following subsections. A part of the experimentation guidelines mentioned, goals are defined taking into account the Goals/Question/Metrics (GQM) approach (Basili et al., 1994).

3.1. Goals

The two research goals of the paper presented in Table 2 below were defined according to the template suggested in Wohlin et al. (2012).

For investigating G1, a business process models from two different real-life information systems retrieved by using reverse engineering techniques were used first. Then the same models were refactored and used in the experiment.

For both quality features, i.e., understandability and modifiability, gain in effectiveness was computed by comparing the correctness and precision of the answers provided to the evaluation questionnaire

Table 2
Research goals.

ID goal	Description
G1	Analyze refactored and non-refactored business process models with the purpose of evaluating them with respect to their understandability and modifiability, from the point of view of the researchers, in the context of a university course in software engineering with bachelor students
G2	Analyze artifact-based and human-perceived measures collected with the purpose of evaluating their relationships with the understandability and modifiability, from the point of view of the researchers, in the context of a university course in software engineering with bachelor students.

Table 3
Description of experimental material.

Real-world system	ID material	Size	Connectivity	Density	Separability	Depth
Tabula	M1$_0$	18	0.889	0.105	15	23
	M1$_R$	18	1.667	0.196	13	75
	M2$_0$	25	0.680	0.057	23	26
	M2$_R$	21	1.190	0.119	19	62
XCare	M3$_0$	57	0.614	0.022	46	203
	M3$_R$	19	1.105	0.123	13	65
	M4$_0$	57	1.158	0.041	36	556
	M4$_R$	20	1.400	0.147	13	136
	M5$_0$	82	1.134	0.028	50	525
	M5$_R$	25	1.320	0.110	19	81

designed. Efficiency, on the other hand, was assessed by focusing on the time spent understanding and modifying the models in both case (reverse and refactored).

For what concern G2, the investigation was aimed to check whether understandability and modifiability data collected according to the artifact-based measures available in literature, are related to the understandability and modifiability perceived by humans.

This goal helps researchers to investigate common wisdom such as *'a business process model of a smaller size is more understandable and modifiable'*.

3.2. Experimental subjects

The experimental units of the experiment (i.e., participants) are 65 Computer Science Bachelor's degree students at the University of Bari, who were enrolled in the subject of Software Engineering. This subject is taught in the second year, over a total of three years, as set out in the B.*Sc*. syllabus of the university in question. The sampling strategy used was "by convenience" (Marshall, 1996). A *convenience sample* is a type of non-probability sampling method where the sample is taken from a group of people easy to contact or to reach. This type of sampling is also known as grab sampling or availability sampling. There are no other criteria to the sampling method except that people be available and willing to participate. In addition, this type of sampling method does not require that a simple random sample is generated, since the only criteria is whether the participants agree to participate. Although this is not a rigorous technique (it considers the selection of the most accessible subjects), it is the least costly to the researcher, in terms of time, effort and money.

The participants were divided into two groups, taking into account their experience and skills, to avoid their different ranges of levels of experience influencing the outcome of the experiment (experience block design). A pre-test (M. Fernández-Ropero et al., 2013) was used to acquire preliminary knowledge about the students and thus attain two balanced and heterogeneous groups. The pre-test collected information about each student, such as his/her attendance ratio (A), his/her average academic grades (M), and his/her skills as regards BPMN (B) and other related notations such as UML (U) and Petri-Net (P). The attendance ratio (*A*) was provided by the course lecturer and corresponded to the rate of attendance on the course. It is important to note that the Italian academic average is a value from 18 to 30, with 30 being the highest grade. Other skills (U, B and P) were evaluated with values 0 to 4, corresponding with the following flags: very poor, poor, average, good and very good. This information was then normalized and used to rank students, such that the experience level of the groups was balanced. The formula used to rank student is shown in (1). Having ranked list of students, the groups were conformed getting one by one in alternative groups.

$$rank = 0.2*A + 0.5*M/6 + 0.1*U/4 + 0.1*B/4 + 0.1*P/4 \tag{1}$$

It is worth noting that the students were not graded on their performance during the course of the experiment in order to avoid social threats originating from evaluation apprehension. Student participation was, in any case, motivated by means of their receiving extra points in their final evaluation at the end of the course.

The reason for using students as experimental subjects rather than real-life software practitioners is that it is possible to have a larger number of subjects. What is more, using students as experimental subjects ensures that their prior knowledge is fairly homogeneous, thus providing an opportunity to test initial hypotheses (Sjøberg et al., 2005). Nevertheless, an experiment with practitioners is also required to reinforce the results obtained (Cardoso, 2006).

3.3. Experimental units

Bearing the goals to be attained in mind, the experimental objects consist of five business process models (see Table 3), which were evaluated under two treatments. The first treatment was that of using original business process models automatically retrieved by means of reverse engineering from two different real-life information systems. The second treatment consisted of using the same sample of retrieved business process models after the application of refactoring (more details about this are provided in the rest of the paper). The ID of each piece of experimental material follows the notation Mi_T, where i is the business process model (1 to 5) and sub-index T is the treatment (0 for original model and R for refactored model).

Fig. 1 shows the process flow followed to obtain the experimental material. The starting point is the "*Tabula*" and "*Xcare*" legacy information systems. Business process models are then mined from the source code using MARBLE (Modernization Approach for Recovering Business process from LEgacy systems) (R. Pérez-Castillo et al., 2011). MARBLE represents business processes according BPMN (OMG, 2011), a well-known graphical notation whose objective is to be easily understood by both system and business analysts. This business process archeology tool was chosen because it is able to retrieve business process models from Java-based source code by means of reverse engineering. MARBLE is released as an Eclipse plug-in and can be easily integrated with other tools.

These source business process models (Model M1, M2, M3, M4 and M5) were retrieved from two real-life information systems. More than one subject system was included in the experimental design, since this might lead to significantly different results. In this study we opted for the use of real-world systems rather than toy systems, since small systems make it difficult to generalize the results. Models M1 and M2 were obtained from the information system belonging to *Tabula,* a web application of 33.3 KLOC (thousands of lines of code) with which to create, manage and simulate decision tables in order to associate conditions with domain-specific actions. Models M3, M4 and M5 were obtained from the *XCare* information system, a mobile application of

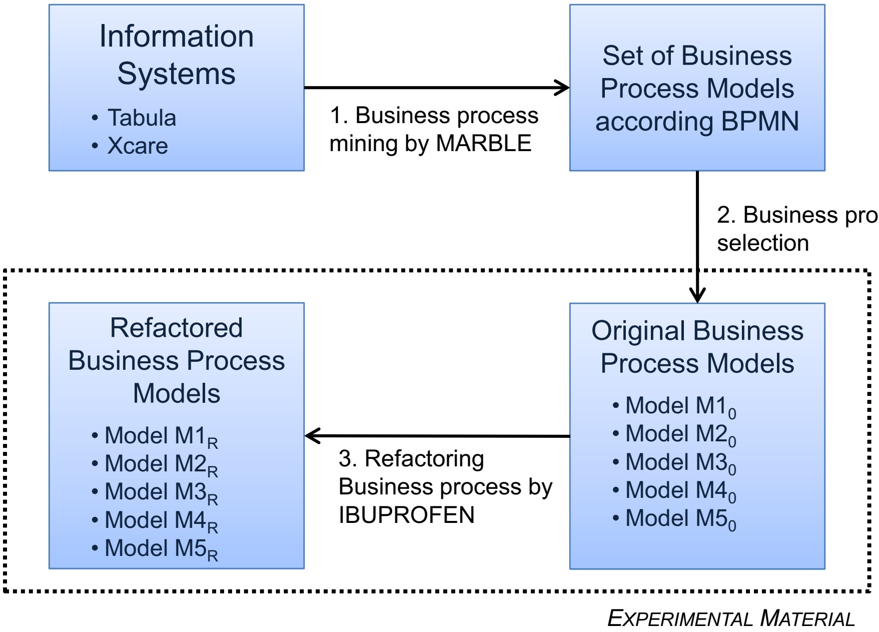

Information Systems
• Tabula
• Xcare
1. Business process mining by MARBLE
Set of Business Process Models according BPMN
2. Business process selection
Original Business Process Models
• Model M1$_0$
• Model M2$_0$
• Model M3$_0$
• Model M4$_0$
• Model M5$_0$
3. Refactoring Business process by IBUPROFEN
Refactored Business Process Models
• Model M1$_R$
• Model M2$_R$
• Model M3$_R$
• Model M4$_R$
• Model M5$_R$
EXPERIMENTAL MATERIAL

Fig. 1. Process flow for obtaining the experimental material.

9.9 KLOC intended for diabetes patients, which analyzes blood (using an external device) and suggests diet plans.

Both systems were selected from Italian companies which were customers of SER&Practices, which is a company that was born as a spinoff from SERLab at University of Bari where the experiment takes place. Due to disclosure agreement signed with SER&Practices and its customers, source code cannot be publicly exposed. Since the business process models were obtained from Italian information systems, it was supposed that the labels of their elements would be well understood by the participants. The experimental material was consequently not translated. An example of a business process model used in the experiment is presented in Appendix II-Fig. 9.

A refactoring technique was then used to refactor the five models. The technique chosen was IBUPROFEN (M. Fernández-Ropero et al., 2013) (Improvement and BUsiness Process Refactoring OF Embedded Noise), a framework with which to refactor business process models specifically retrieved from existing information systems. The technique allows different refactoring operators to be applied, which are grouped into three categories according to their behavior: (i) maximization of relevant element; (ii) fine-grained granularity reduction; and (iii) completeness maximization. We should add that IBUPROFEN has been implemented as an Eclipse plug-in, and it was specially designed for business process models represented according to the BPMN. It is thus easy to use IBUPROFEN in combination with MARBLE. The process of refactoring was conducted by a domain and BPMN expert in order to obtain the refactored business process models employed in the experiment. Table 3 shows the values for the artifact-based measures used to assess the understandability and modifiability of the experimental material before and after applying refactoring (see Appendix I - Table 16): *Size* is the number of nodes in a business process model (i.e., business tasks, gateways, data objects and events); *Connectivity* is, in turn, the ratio between the total number of arcs in a business process model (i.e., sequence flows and associations) and the total number of nodes; *Separability* represents the ratio between the number of cut-vertices in a business process model (i.e., nodes that serve as bridges between otherwise strongly-connected components) and the total number of nodes in the business process model; *Density* is the ratio between the total number of arcs in a business process model and the theoretical maximum number of possible arcs regarding the number of nodes; *Depth* defines the maximum nesting of structured blocks in a process model. This measure affects the understandability of a business process model.

Table 3 also shows that Model M1 maintains the same number of elements (*Size*) but that the connectivity between them has increased. The depth of that model is greater after refactoring than the original depth, while the density is slightly greater when refactoring is applied. In the case of the size variation throughout all the models, the size of refactored models is less than, or equal to, the size of non-refactored ones. With regard to the connectivity, there is an increase in this variable when refactoring is applied. It is interesting to note that the difference between connectivity before and after refactoring is lower when the model is larger. The same occurs with density; refactored models are denser than non-refactored ones. However, the difference in density is greater when the model is larger. The separability of refactored models is lower than the separability of non-refactored ones. With regard to the depth of refactored and original models, this measure tends to be greater in M1 and M2 when refactoring is applied. However, the depth of the other models (M3, M4 and M5) tends to be lower after refactoring. This means that the behavior of refactoring as regards depth is different depending on the model size.

3.4. Tasks

A set of relevant tasks was defined for each business process model in the experimental material (see Table 4). The set of tasks was designed in such a way that very basic or trivial tasks were not included, while the tasks chosen were close in the scope and complexity of the real tasks carried out by practitioners.

The participants were therefore asked to complete two parts for each piece of experimental material in order to evaluate both effectiveness and efficiency (see Table 4):

- *Understandability part*: The participants were requested to fill in a form so as to assess their capability of understanding the business process model being studied from two perspectives:
  - ○ artifact perspective (U_{art}). This questionnaire consisted of 6 questions about the business process model to evaluate its understandability: 5 true-or-false questions and 1 open-ended question. The time spent on answering this part was registered to evaluate efficiency. An example is shown in Appendix II.

Table 4
Summary of tasks for each piece of experimental material.

Part	ID task	Task	Time	# Questions	Type answer
Understandability part	U_{art}	artifact based	Yes	5	True-or-false
				1	Open-ended question
	U_{hum}	human based	No	5	Seven-point scale
Modifiability part	M_{art}	artifact based	Yes	2	Multi-choice questions Open-ended question
	M_{hum}	human based	No	5	Seven-point scale

- ○ human perspective (U_{hum}). To obtain more faithful results, this part provided a set of 5 subjective questions (see Appendix II). The first question asked about difficulty, i.e., how the experimental material hindered answering the questionnaire. The possible answer was defined by means of a seven-point scale in which "1" was very easy and "7" was extremely difficult. The remaining questions asked how certain quality faults negatively affected the understandability of the model. These quality faults were: isolated nodes, missing gateways in branches, bidirectional flows and missing starting and ending business activities. The possible answers were similarly defined using a seven-point scale, in which "1" meant that understandability had not been affected and "7" meant that it had been greatly affected.

- *Modifiability part*: The participants were requested to fill in a form to assess their capability of making modifications to the business process model being studied, from an objective and a subjective perspective:
  - ○ artifact perspective (M_{art}). This questionnaire requested subject to modify some aspects of the business process model. It was composed of two kinds of questions: multi-choice questions and/or open-ended questions. The time spent answering this part was also recorded to evaluate the efficiency. An example is shown in Appendix II.
  - ○ human perspective (M_{hum}). In a similar way to the understandability part, this section provided a set of 5 subjective questions. The intention of the first of these was to specify the level of difficulty involved in modifying the model. The remaining questions asked about how the aforementioned quality faults harmfully affected the modifiability of the business process models. All the questions were quantified using a seven-point scale. These questions were similar to the tasks in the previous part, but were related to modifiability rather than understandability.

Despite the fact that the time needed to solve tasks U_{art} and M_{art} was recorded, no time limit per task was imposed in order to avoid inaccurate answers that could have been the result of the participants being placed under further pressure.

The human-perceived tasks in both parts (U_{hum} and M_{hum}) were always the same in each treatment; the sole difference lay in the experimental material provided to the subjects. In the case of artifact-based tasks, these tasks had to be specially defined for the specific experimental material provided in each treatment, although all the tasks had the same level of complexity.

At the end of the experiment, the subjects filled in a post-test whose aim was to provide further feedback on the conduction of the experiment. This post-test asked the subjects about the difficulty involved in completing the tasks, along with the overall time pressure. The questionnaire followed the five-point *Likert* scale to define the answers to each question (Oppenheim, 2000). This feedback is useful as regards improving the design and conduction of future replications. This post-test is available online in M. Fernández-Ropero et al. (2013).

The material was originally written in English, but owing to the fact that the participants in the experiment are Italian speakers, all of the questionnaires (pre-test, understandability part, modifiability part and post-test) were translated into Italian to make them easier to understand. Furthermore, open-ended questions were specially defined to allow the subjects to complete tasks by adding some sketches; all of this made it easier to evaluate the tasks.

Since the business process models of the experimental material followed the BPMN notation, the subjects were provided with a leaflet in Italian containing all the basic BPMN elements. This leaflet was provided to mitigate the threat derived from a poor awareness of BPMN on the part of the subjects.

3.5. Hypotheses and variables

This subsection shows the hypotheses formulated to address the proposed goal. In the case of both research goals (see Table 2), the null hypothesis, denoted as H_{0ij}, and its corresponding alternative hypothesis, denoted as H_{1ij}, need to be formally described, where *i* corresponds to the goal identifier, and *j* is a counter wherever more than one hypothesis is formulated per goal.

H_{011} *: There is not significant difference in understandability between refactored and non-refactored business process models.*
H_{111}*:* H_{011}
H_{012}*: There is not significant difference in modifiability between refactored and non-refactored business process models.*
H_{112}*:* H_{012}
H_{013}*: There is not significant difference in understandability efficiency between refactored and non-refactored business process models.*
H_{113}*:* H_{013}
H_{014}*: There is not significant difference in modifiability efficiency between refactored and non-refactored business process models.*
H_{114}*:* H_{014}
H_{021}*: There is no correlation between artifact-based Understandability and human- perceived ones.*
H_{121}*:* H_{021}
H_{022}*: There is no correlation between artifact-based Modifiability and human- perceived ones*
H_{122}*:* H_{022}

The *independent variable* employed to answer H_{x1j}(where *x* is null or alternative and *j* is the number of hypothesis) is the treatment used (*T*), i.e., with and without refactoring.

The *dependent variables* are defined by means of the following measures:

- *artifact-based Understandability Effectiveness (UEffec)*: This measure was defined as the number of correct answers obtained for the U_{art} tasks out of the total number of questions for the U_{art} tasks. The value of UEffec is between 0 (none correct) and 1 (all correct).
- *human-perceived understandability Effectiveness (uEffec)*: This measure is defined as the normalized value of how the participants perceived understandability. This variable name is distinguished from the previous one by using 'u' in lower case rather than 'U' in upper case. The value 0 corresponds to "very difficult to understand", while the value 1 corresponds to "very easy to understand". In addition, the experiment considers 4 metrics in order to evaluate the impact of the quality flaws mentioned on the effectiveness of understandability. All the metrics share the same definition range of values and correspond to the answers obtained for the U_{hum} tasks.
  - ○ *uEffecI*: The value 0 means that isolated and sheet nodes have no effect as regards understanding the business process model, while the value 1 signifies that these quality faults have a negative

effect.
  - *uEffecG*: The value 0 means that missing gateways in branches have no effect as regards understanding the business process model, while the value 1 signifies that these quality faults have a negative effect.
  - *uEffecB*: The value 0 means that bidirectional flows have no effect as regards understanding the business process model, while the value 1 signifies that these quality faults have a negative effect.
  - *uEffecS*: This measure corresponds with how easy it was to identify tasks executed at the beginning and at the end in order to understand the model. The value 0 means "very difficult" while the value 1 means "very easy".
- *artifact-based Modifiability Effectiveness (MEffec)*: This measure is defined as the number of correct answers obtained for the M_{art} tasks out of the total number of questions for the M_{art} tasks. The value of MEffec is between 0 (none correct) and 1 (all correct).
- *human-perceived modifiability Effectiveness (mEffec)*: This measure is defined as the normalized value of how the participants perceived modifiability. This variable name is distinguished from the previous one by using 'm' in lower case rather than 'M' in upper case. The value 0 corresponds to "very difficult to modify", while the value 1 corresponds to very easy to modify. The experiment also considers 4 metrics in order to evaluate the impact of the quality flaws mentioned on the effectiveness of modifiability. All the metrics share the same definition range of values and correspond to the answers obtained for the M_{hum} tasks.
  - *mEffecI*: The value 0 means that isolated and sheet nodes have no effect as regards modifying the business process model, while the value 1 signifies that these quality faults have a negative effect.
  - *mEffecG*: The value 0 means that missing gateways in branches have no effect as regards modifying the business process model, while the value 1 signifies that these quality faults have a negative effect.
  - *mEffecB*: The value 0 means that bidirectional flows have no effect as regards modifying the business process model, while the value 1 signifies that these quality faults have a negative effect.
  - *mEffecS*: This measure corresponds to how easy it was to identify tasks executed at the beginning and at the end in order to modify the model. The value 0 means "very difficult", while the value 1 means "very easy".
- *Understandability Efficiency (UEffic)*: This measure is defined as the time (in seconds) taken to answer the questions for the U_{art} task (related to understanding the model).
- *Modifiability efficiency (MEffic)*: This measure is defined as the time (in seconds) taken to answer the questions for the M_{art} task (related to modifying the model).

In order to address goal G2 corresponding to $H_{021},$ H_{121}(understandability) and $H_{022},$ H_{122}(modifiability), *UEffec* and *MEffec* (artifact-based) were used as *dependent variables* for the correlation analysis. Moreover, in order to evaluate the hypothesized correlation of understandability and modifiability, some metrics for evaluating understandability and modifiability were used as *independent variables* (see Table 5). These selected metrics were the size of the business process model, the connectivity, separability, density and depth, which were defined in Section 3.3.

In both cases, the experiment employs the source model (M1, M2, M3, M4 and M5) as the *moderating variable* (see Table 3). The outcome may be different depending on which particular model is focused upon. Table 5 summarizes the set of variables used in the experiment, along with their abbreviations by which they will be referred throughout the document.

3.6. Experimental design

The experiment is a repeated, *within subjects* design since every subject applies both treatments, refactored/non-refactored, although to different systems). The assignation of experimental units and material to each group was carried out as follows:

Table 5
Variables definition.

Abbreviation	Description
UEffec	artifact-based understandability effectiveness
uEffec	human-perceived understandability effectiveness
uEffecI	Effect on the understandability effectiveness
uEffecG	
uEffecB	
uEffecS	
MEffec	artifact-based modifiability effectiveness
mEffec	human-perceived modifiability effectiveness
mEffecI	effect on the modifiability effectiveness
mEffecG	
mEffecB	
mEffecS	
UEffic	Understandability efficiency
MEffic	Modifiability efficiency
Model	Source BP model
T	Treatment
Size	Size
Conn	Connectivity
Den	Density
Dep	Depth

- The prior experience effect was used to assign subjects to groups. The participants were, therefore, assigned to a group according to the ranking provided by means of the pre-test; this was done to alleviate the prior experience effect and ensure that there were groups with a balanced level of experience.
- Each group was given 5 questionnaires with their respective understandability and modifiability tasks, as mentioned previously. Each questionnaire corresponded to one of the five business process models under only one of the treatments (refactored or non-refactored). If one group of subjects was performing the tasks concerning a business process model (experimental unit) under a treatment, this group did not perform the tasks regarding the same model with the opposite treatment. Table 6 shows the distribution of the experimental material in the two groups. The table shows which business process model under which treatment was covered in each questionnaire, using the notation Mi_T (see Table 3), where *i* is the business process model (1 to 5) and sub-index *T* is the treatment (0 for original model and R for refactored model). This arrangement of experiments is called Latin square (Juristo and Moreno, 2013).

Moreover, in order to mitigate possible side effects related to the participants' different levels of expertise, they were also given a background lecture before the experiment session.

This design was established after analyzing risks of crossover experiments in software engineering as provides by Vegas et al. (2016). In this experiment, the carryover is not considered since the persistence of the effect of one treatment and the application later of the second treatment is not desired.

3.7. Execution procedure

The whole experiment was organized in two different sessions. The background lecture was given in the first session, while the experiment was carried out in the second.

Phase 1. Background session: This phase was, in turn, organized as follows:

1. A *background lecture* was given in an effort to provide detailed instructions about the experiment and the main concepts of business

Table 6
Experimental design. Questionnaires for each group.

Group	ID material Questionnaire 1	ID material Questionnaire 2	ID material Questionnaire 3	ID material Questionnaire 4	ID material Questionnaire 5
Gr1	$M1_0$	$M2_R$	$M3_0$	$M4_R$	$M5_0$
Gr2	$M1_R$	$M2_0$	$M3_R$	$M4_0$	$M5_R$

process and reverse engineering. Details of the experimental hypotheses, along with the two treatments to be compared, were hidden from the students so that the results would not be affected or conditioned.

2. After the background lecture, the subjects carried out a *training example* with tasks that were similar to those in the experiment. This training example was aided by the instructor's explanations and was carried out without time restrictions. Both the background lecture and the training example were carried out a few days before the experiment session so that the subjects could better assimilate the knowledge.
3. At the end of this session, the subjects filled in the *pre-test* (cf. Section 3.2). The pre-test data was then typed and managed on an excel sheet in order to rank all the students and establish the two subject groups.

Phase 2. Experiment session: This phase was, in turn, organized as follows.

1. This session started with a very brief summary of the background lecture, which focused particularly on the BPMN notation.
2. The subjects were then provided with clear instructions on how to conduct the experiment:
   ○ The tasks had to be done in order, one after another.
   ○ It was not possible to return to a previous completed task
   ○ The starting and finishing times had to be written accurately. It was absolutely necessary to record the time from a common clock located in the classroom. A clock application, in which the time interval was 5 seconds rather than 1 s, was specially designed for this experiment. 5 seconds was chosen because this is the estimated time that humans spend recording the time. Possible time accuracy errors were thus mitigated.
3. A similar, but completed questionnaire was also given to the subjects to ensure that they completed all the experiment tasks accurately. The sample questionnaire is provided in Appendix II and contains experimental material (a business process model of similar complexity) and questionnaires with very similar tasks to those related to the understandability and modifiability parts. The running example was carried out by the subjects in a simulation, with the instructor's help.
4. The experiment was then conducted in the classroom, where the students were supervised by the instructor and were not able to communicate with each other. All the material was distributed to them according to the group to which they were assigned and taking into account the pre-test information. They received the experimental material (business process models) first, in order, and were then provided with the tasks, which had to be completed in the same order.
5. At the end of the experiment, the subjects were encouraged to complete the post-test, with which feedback about the conduction of the experiment could be obtained. This had no time limit.

Phase 3. Post-execution activity: This phase was organized as follows.

1. Firstly, data collected from each subject was typed at the end in an SPSS file so that it could be appropriately managed and analyzed.
2. The analysis procedure (which is set out in the following section) was then carried out to evaluate the research questions that had been established.

3.8. Analysis procedure

The analysis procedure consisted of two types of analyses. These analyses were performed for each research goal:

1. Descriptive statistics: In this step, the data were described, analyzed and represented using numerical and graphical methods in order to summarize and present the information contained in them. The main features of data collection were therefore quantitatively described. The mean was used to describe the central tendency of the data set, while the standard deviation (or variance) was used to describe the variability or dispersion of the sample.
2. Statistical hypothesis testing: In this step, the data were analyzed in order to reject or not reject some a priori assertions (called hypotheses). There are parametric and non-parametric tests, depending on whether or not the sample is normally distributed. For the first goal, univariant tests were used to compare the results obtained for a refactored and non-refactored business process model from the controlled experiment. For the second goal, correlation tests were used to measure the statistical dependence between the hypothesized and perceived measures. The significance level selected for both tests was 0.05, which corresponds to a confidence level of 95%.

A normality test was first carried out to choose the most suitable tests. The Shapiro-Wilk test was thereby applied to check the normality of the *dependent variables* across levels of *independent variables* (cf. Appendix III). The test verified that the sample did not have a normal distribution, since the null hypothesis was rejected. However, when the division by experimental material was performed, the distribution was normal in some cases. It was not, however, possible to apply parametric tests in this experiment.

With regard to the tests, the non-parametric test chosen to check H_{x1j}(where *x* is null or alternative and *j* is the number of hypothesis) and to compare data in the different treatments (corresponding to G1) was the *Mann-Whitney U test*. This test was chosen because the *independent variable* (*T*) had only two possible values (R and 0), the *dependent variables* were quantitative, and there was no relationship between the samples of groups. This test checked whether there was a significant difference between the *dependent variables* with regard to the *independent variable*. In this test, the null hypothesis was that two populations (according to *T*) are the same, as opposed to an alternative hypothesis, which was that one particular population would tend to have larger values than the other. The hypotheses for this test were, therefore:

H_0: There is no significant difference between refactoring (T = R) and without refactoring (T = 0)
H_1: There is significant difference between both treatments.

The non-parametric test chosen to check H_{x2j}(where *x* is null or alternative and *j* is the number of hypothesis) in order to discover the correlation between each pair of variables (corresponding to G2) was, meanwhile, *Spearman's correlation test*. Spearman's rho (ρ) is the degree to which the real values of the *dependent variable* are close to the predicted values (*independent variables*); it is between −1 and 1.

Table 7
Statistical tests (x is null or alternative; U/M are artifact-based, u/m are human-perceived).

Hypothesis	Distribution	Independent variable	Dependent variables	Relationship between samples	Statistical test
H_{x11}	No normal	*T* (treatment)	*UEffec* and *uEffec*	No relation	Mann-Whitney U test
H_{x12}			*MEffec* and *mEffec*		
H_{x13}			*UEffic*		
H_{x14}			*MEffic*		
H_{x21}		*Size, Connectivity, Separability, Density* and *Depth*	*UEffec* and *uEffec*		Spearman's correlation test
H_{x22}		*Connectivity, Density* and *Separability*	*MEffec* and *mEffec*		

When both variables are perfectly monotonically related, the coefficient becomes 1. The sign of the Spearman correlation indicates the direction of association between X (*independent variable*) and Y (*dependent variable*):

- $\rho > 0$: Y tends to increase when X increases.
- $\rho < 0$: Y tends to decrease when X increases.
- $\rho = 0$: there is no tendency for Y to either increase or decrease when X increases.

Table 7 shows the statistical tests applied in the experiment according to the proposed hypotheses to be tested (where x is null or alternative), the distribution and the dependent and independent variables in each one, and the relationship between samples.

4. Execution

This section shows how the experimental procedure was enacted. The procedure is explained step by step in Section 4.1, while Section 4.2 shows the deviations from the plan that occurred.

4.1. Preparation

The experimental sessions took place on two different days (3 days apart). The first session lasted two hours. Of these two hours, 70 minutes were spent explaining the background (introduction, business process concepts and BPMN notation) and 15 minutes were spent on the training example, with a BPMN business process model. After a break of 20 minutes, the last part of the background lecture regarding the reverse engineering of business process models was taught for the last 15 minutes.

Although 80 students had normally attended lectures on the subject of *Software Engineering,* the first session took place with 68 students, while the second session (the conduction of the experiment) took place with a final total of 65 subjects.

After the first session, the pre-test data were analyzed and scored in a ranking. The students were assigned to one of two groups, as prescribed by the experimental design. The number of students was balanced, as was the mean score, as Table 8 shows.

The second session took two and a half hours. The first 30 minutes were spent performing the running example with similar tasks. In the remaining two hours, the subjects carried out the experiment tasks.

Three students who had not attended the first session, and had not therefore completed the pre-test, were assigned to the subject groups randomly. We attempted to achieve groups of the same size.

Table 8
Distribution of students after session 1 and session 2.

	Session 1: Background lecture		Session 2: Experiment conduction	
	#Students	Mean score (pre-test)	#Students	Mean score (pre-test)
Gr1	34	0.5049	34	0.5049
Gr2	34	0.5071	31	0.4882

4.2. Deviations

This subsection discusses the problems detected during the execution of the experiment, and how they were fixed.

After performing the experiment, data from the experiment and post-test were typed in an *SPSS* file to be analyzed. During this activity, the following data quality faults were detected:

- *Absenteeism*: Three students that performed the pre-test, and who were assigned to a group, did not attend the second session.
- *Missing recorded time*: Some students forgot to record the time before and/or after doing the tasks.
- *Missing answers*: Some questions were not answered and were left blank.

Those students who did not attend the second session were removed from the study. Unfortunately, these three students had been assigned to the same group (Gr2) and their absence meant that the number of participants in each group was not balanced, as Table 8 shows.

The fact that the time was not recorded for some tasks prevented us from calculating *UEffic* and *MEffic*. The times for a total of 8 tasks were wrongly recorded by the subjects. The values of these variables were not taken into account during analysis.

Missing answers in objective tasks were considered as a failure. In the case of missing answers in subjective questions, three different actions were triggered:

- For the first and last questions related to the effect of identifying tasks executed at the beginning or at the end of the business process model: No action was taken in this case, since it was considered as absenteeism; these are considered as missing values.
- The remaining questions related to the effect of isolated nodes, missing gateways in branches and bidirectional flows: a blank reply was considered as 0, since the zero value was allowed in this case; this indicates that the quality faults mentioned have no effect on the understandability or modifiability.
- On the other hand, blank answers in the post-test were not specifically addressed, since they did not have any influence on the results of the experiment.

5. Analysis

This section reports the results of the experiment after the data analysis. The data set preparation phase is addressed first. Descriptive statistics are then shown and explained. Finally, the hypotheses formulated are tested individually, to achieve a more accurate and in-depth analysis.

5.1. Data set preparation

Dependent variable data were transformed as follows so as to normalize their values between 0 and 1 (see Table 5).

- *uEffec* and *uEffecS*: Since the answers to the questionnaire followed a seven-point scale, the answer a was transformed by following formula (2). This is owing to the fact that the original scale was between 1 and 7 (see Section 4.2):

$$uEffec = 1 - \frac{a-1}{6} \tag{2}$$

- *uEffecI*, uEffecG and *uEffecB*: Since the answers to the questionnaires followed a seven-point scale, the answer a was transformed by following formula (3). This is owing to the fact that the original scale was between 0 and 7:

$$uEffecI = 1 - \frac{a}{7} \tag{3}$$

- Dependent variables related to modifiability underwent the same transformations as those related to understandability, i.e., *mEffec, mEffecS* were transformed according to (2); and *mEffecI, mEffecG* and *mEffecB* were adapted using (3).

5.2. Descriptive statistics

Although the whole collected data set is available online in M. Fernández-Ropero et al. (2013), information concerning the pre-test is assessed first. Table 8 provides information about the number of participants that responded to the questionnaires in each group. Fig. 2 presents some statistics concerning the participants' attendance ratio, average academic grades and skills. Fig. 2 provides data from the pre-test of 62 subjects, (this smaller number is owing to the absenteeism of three subjects). These statistics show that most of the subjects attended the course regularly (over 75% of the lectures). Moreover, the academic grades are mostly distributed among Normantas and Vasilecas (2013), Canfora et al. (2011), M. Fernández-Ropero et al. (2013), Dumas et al. (2011), Ekanayake et al. (2012), Leopold et al. (2012), Pittke et al. (2013). Furthermore, the participants' skills were very poor as regards Petri-Nets. UML was the best- known notation, while the skills related to BPMN were poor or very poor.

The following subsections show the descriptive statistics for the experimental data. Descriptive statistics regarding effectiveness- and efficiency-related variables are set out in the first subsection (corresponding to experimental goal G1), while the descriptive statistics concerning hypothesized and perceived understandability/ modifiability are presented in the second subsection (corresponding to experimental goal G2).

5.2.1. G1. Effectiveness and efficiency statistics

Both Table 9 and Table 10 show the main descriptive statistics regarding the effectiveness- and efficiency-related variables, respectively. For each dependent variable these tables provide: the number of cases (*N*) for each treatment (*T*); its mean ($\overline{X}$), and its standard deviation (*SD*). Table 10 represents time in seconds.

We should state that effectiveness is greater after refactoring for every dependent variable (see Table 9). Another insight obtained upon observing Table 9 is that the mean difference for objective tasks (*UEffec* and *MEffec* as artifact-based) is greater than the mean difference for subjective ones (*uEffec* and *mEffec* as human-perceived).

With regard to efficiency, the average time needed to perform all the different tasks (both understandability- and modifiability-related) with refactored business process models was less than that spent understanding or modifying non-refactored models. In fact, the understanding of refactored models was, by and large, 24% faster than the understanding of original models. In addition, the refactored models were, on average, modified 115% faster than the original ones (see Table 10).

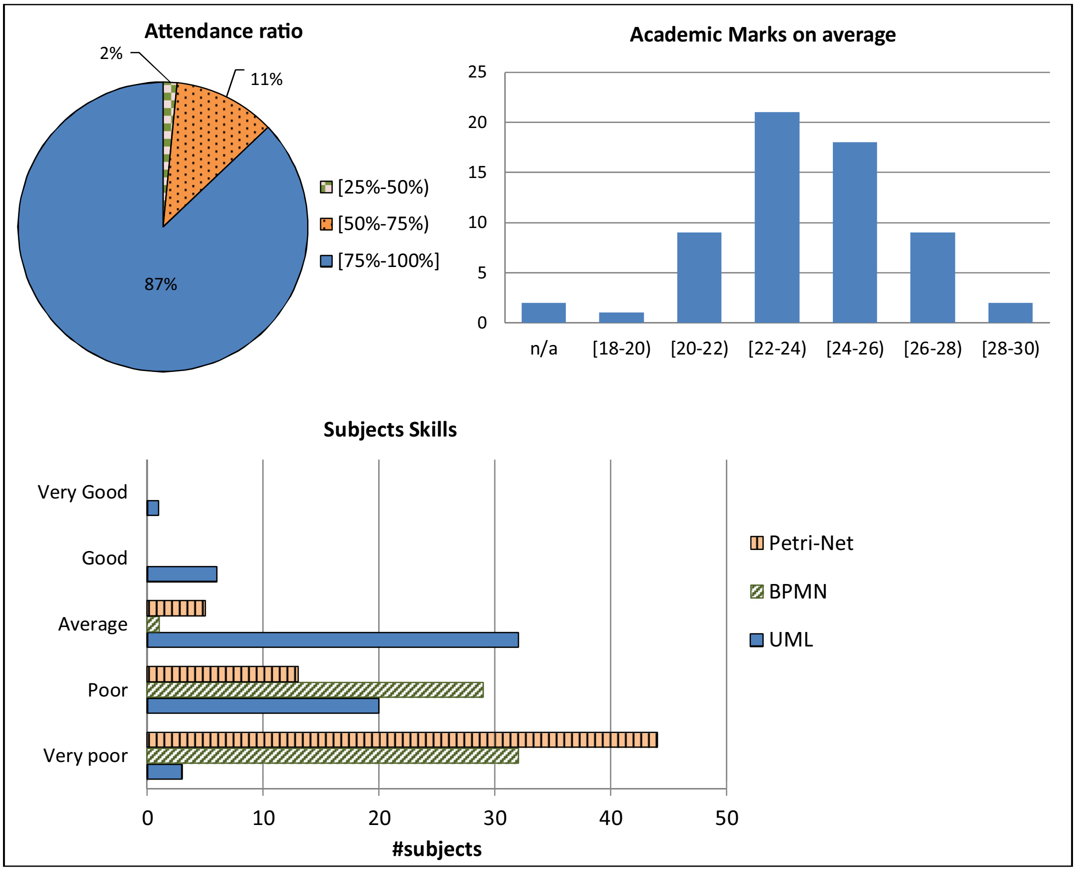


Fig. 2. Statistics of pre-test (after experiment).

Table 9
Descriptive statistics for effectiveness-related variables for all and for each source model (U/M are artifact-based and u/m are human-perceived).

Model	T	UEffec			uEffec			MEffec			mEffec		
		N	$\overline{X}$	*SD*	*N*	$\overline{X}$	*SD*	*N*	$\overline{X}$	*SD*	*N*	$\overline{X}$	*SD*
All	0	164	0.7226	0.2187	162	0.5453	0.3030	164	0.5335	0.3509	162	0.6183	0.2797
	R	161	0.9068	0.1316	160	0.7354	0.2490	161	0.8975	0.2174	161	0.7153	0.2692
M1	0	34	0.9216	0.1435	33	0.8081	0.2504	34	0.5441	0.2572	34	0.7647	0.2500
	R	31	0.9409	0.0918	31	0.7366	0.2645	31	0.9677	0.1249	31	0.9111	0.2559
M2	0	31	0.7419	0.7419	31	0.5484	0.2516	31	0.5806	0.3674	31	0.6237	0.2356
	R	34	0.8676	0.1730	34	0.2516	0.2390	34	0.8088	0.2756	34	0.6667	0.2496
M3	0	34	0.8529	0.1282	34	0.4853	0.2334	34	0.6471	0.3800	34	0.5490	0.2480
	R	31	0.9624	0.9624	31	0.7688	0.2094	31	0.9677	0.1249	31	0.7043	0.2182
M4	0	31	0.5068	0.5068	31	0.4355	0.3002	31	0.3710	0.4076	31	0.5323	0.3116
	R	34	0.8971	0.1422	34	0.8235	0.2460	34	0.9412	0.1635	32	0.6961	0.2972
M5	0	34	0.5049	0.5049	33	0.4444	0.3191	34	0.5147	0.2883	32	0.6146	0.3006
	R	31	0.8710	0.1270	30	0.6667	0.2626	31	0.2883	0.2792	31	0.6989	0.3056

Table 10
Descriptive statistics for efficiency-related variables for all and for each source model (seconds).

Model	T	UEffic			MEffic		
		N	$\overline{X}$	*SD*	*N*	$\overline{X}$	*SD*
All	0	164	233	134.1445	161	184	86.3489
	R	160	187	84.0575	159	154	86.648
M1	0	34	85	44.4666	34	159	81.0470
	R	31	138	81.1198	31	83	37.5925
M2	0	31	242	101.2678	31	165	80.6167
	R	33	230	101.4505	33	140	85.9751
M3	0	34	299	62.4830	33	243	103.2516
	R	31	204	73.5962	30	226	80.1298
M4	0	31	323	192.0698	30	200	70.7357
	R	34	178	74.2176	34	197	87.0953
M5	0	34	225	74.5591	33	154	59.4387
	R	31	186	58.8359	31	120	42.4788

Although the overall descriptive statistics are shown in the above tables, the source model is a moderating variable that may provide different results. The descriptive statistics for each source model are, therefore, illustrated separately in Table 9 and Table 10 to allow more a precise and in-depth analysis, in addition to strengthening the results.

According to Table 9, the mean of the variables is, in most cases, greater when refactoring is applied. However, with regard to $M1_R$, the mean of *uEffec* is slightly lower than the value of the same measure for $M1_0$. This means that, despite the fact that the participants answered the questions more accurately in the refactored model than in the original one, M1 after refactoring was a little less objectively understandable than before refactoring. Even so, the modifiability of M1 was far greater after refactoring than before it. This therefore ensures that M1 was previously well understood, since the understandability part was carried out first.

With regard to M2, the refactored material was subjectively less understandable than the original material but, as before, there were more correct answers in the understandability and modifiability part when the model had been refactored than when it had not.

With regard to M5, the subjects attained fewer correct answers in the modifiability part (*MEffect*) with refactored models than with the original model. Surprisingly, the subjects modified the refactored model more easily than they did the non-refactored model, as *mEffect* shows.

It should be noted that the number of cases (*N*) in each case varies slightly. The reason for this is that some students left some tasks blank. For example, the subjective understandability effectiveness (*uEffec*) of M1 has 33 rather than the 34 expected cases, since one participant did not fill in this part of the questionnaire.

Table 10 shows that, for all the models, the time spent understanding and modifying refactored business process models is less than that spent understanding and modifying the original business process models. However, the time with regard to *UEffic* in $M1_0$ is lower than in $M1_R$. Although the time difference is high in all cases, *MEffic* in $M4_R$ is just slightly lower than *MEffic* in $M4_0$ (only 2.5 seconds).

Although the descriptive statistics of the aforementioned measures indicate that understandability and modifiability are more effective under treatment with refactoring (*R*), some quality faults of business process models are also analyzed in an effort to discover exactly why this improvement may have occurred. The quality flaws analyzed are those set out in Section 3.5: *xEffecI* (isolated nodes), *xEffecG* (missing gateways), *xEffecB* (bidirectional flows) and *xEffecS* (missing start/end tasks), where *x* can be *u* (related to understandability) or *m* (related to modifiability). Taking into account the changes made and detailed in Section 5.1, a zero value for a quality fault (isolated and sheet nodes, missing gateways, etc.) means that this quality fault has an extremely negative effect on understandability or modifiability, while a value of 1 means the negative effect is low.

Table 11 and Table 12 depict the descriptive statistics of these variables. In all cases the mean is lower when the model has not been refactored than when it has been. This means that these quality faults (isolated and sheet nodes, missing gateways, etc.) have a more negative effect on the understandability and modifiability in the original models. As a result, we can draw the conclusion that these quality faults would seem to be found more commonly in this kind of business process models, without refactoring.

The descriptive statistics in Table 9 and Table 10 could be used to predict that the understandability and modifiability of refactored business process models are more effective and efficient than the understandability and modifiability of the original models. Although this may be the case, Section 5.3.1 strengthens the analysis of the results by

Table 11
Descriptive statistics for negative effect on understandability.

T	uEffecI			uEffecG			uEffecB			uEffecS		
	N	$\overline{X}$	*SD*	*N*	$\overline{X}$	*SD*	*N*	$\overline{X}$	*SD*	*N*	$\overline{X}$	*SD*
0	164	0.5078	0.2814	164	0.5767	0.2943	164	0.5828	0.2512	164	0.4917	0.3599
R	161	0.7276	0.2075	161	0.6974	0.2163	161	0.6886	0.2292	161	0.7990	0.2445

Table 12
. Descriptive statistics for negative effect on modifiability.

T	mEffecI			mEffecG			mEffecB			mEffecS		
	N	$\overline{X}$	*SD*	*N*	$\overline{X}$	*SD*	*N*	$\overline{X}$	*SD*	*N*	$\overline{X}$	*SD*
0	164	0.6246	0.2533	162	0.6071	0.2664	164	0.5993	0.2590	162	0.6232	0.3362
R	161	0.7169	0.2216	160	0.6717	0.2509	161	0.7028	0.2137	161	0.7820	0.2625

Table 13
Hypothesized more understandable/modifiable model, according to the increase/decrease of measures.

Model	Size		Connectivity		Density		Separability		Depth	
	U (-)	M	U (-)	M (-)	U (-)	M (-)	U (+)	M (+)	U (-)	M
M1	R	No effect	0	0	0	0	0	0	0	No effect
M2	R		0	0	0	0	0	0	0	
M3	R		0	0	0	0	0	0	R	
M4	R		0	0	0	0	0	0	R	
M5	R		0	0	0	0	0	0	R	

means of statistical tests, with the objective of verifying the hypotheses formulated at the beginning of the experiment.

5.2.2. G2. Correlation between artifact-based and human-perceived results

The descriptive statistics concerning artifact-based understandability and modifiability were presented in Table 3, in which the values for size, connectivity, density, separability and depth were detailed (cf. Section 3.3). According to the assumptions made in previous works (L. Sánchez-González et al., 2010; Mendling et al., 2007), the effect of each measure affects understandability and modifiability (see Table 13) positively or negatively. Size affects understandability (U) negatively (-); i.e., greater size makes it more difficult to understand a certain business process model. Connectivity also affects understandability and modifiability (M) negatively. This means that lower connectivity values imply that business process models are more understandable and modifiable, owing to a lower level of intricacy. Separability, on the other hand, affects modifiability and understandability positively (+), since a lower separability implies hard and error-prone modifications of business process models. Density affects understandability and modifiability negatively. The lower the density, the more understandable and modifiable the business process model is. Depth has a negative effect on understandability. The lower the depth, the more understandable the business process model is.

In accordance with the above assumptions, Table 13 shows which treatment provides most understandability and modifiability, according to the variation of the measures: size, connectivity, density, separability and depth. For example, as regards M1, the model is more understandable after refactoring than before it, since the model size has decreased. However, as regards the remaining measures, M1$_0$ is more understandable and modifiable than M1$_R$. Table 13 also shows that M2$_0$ is more understandable and modifiable than M2$_R$, except in the case of size. Nevertheless, Table 13 suggests that M3$_R$, M4$_R$ and M5$_R$ are more understandable than M3$_0$, M4$_0$ and M5$_{0,}$ since the depth has been decreased.

With regard to perceived understandability and modifiability, descriptive statistics were presented in Table 9, in which values for effectiveness-related variables were provided. In this case, refactoring was always the best option for all models, thus contradicting Table 13.

However, any conclusions regarding the correlation between these two measurements (artifact-based and human-perceived) can be drawn by observing both tables, since they contain sufficient information. As a consequence, correlation tests are performed in Section 5.3.2 to check the correlation between both measures.

5.3. Hypothesis testing

The data analysis performed to answer each hypothesis is presented in the following subsections. The hypothesis testing is divided into 2 subsections: the first sub-section addresses the effectiveness and efficiency concerns, as expressed in the research goal G1, while the second analyzes the correlation between artifact-based understandability/modifiability values and human-perceived, in relation to G2.

5.3.1. G1. Effectiveness and efficiency

The first goal, G1, is first related to an assessment of how refactoring affects the effectiveness of understanding and modifying business process models. To achieve this goal, a Mann-Whitney test was performed for each of the hypotheses formulated (cf. Section 3.8). The hypotheses for both tests are the following:

H_0: There is not statistically significant difference between groups
H_1: There is a statistically significant difference between groups

Table 14 shows the result of the Mann-Whitney tests. The significance level (*sig*) is showed. All the hypotheses are rejected when all experimental material is considered; since the value of its *p-value* is lower than 0.05 for every dependent variable. This means that there is a significant difference between the values of *UEffec, uEffec, MEffec and*

Table 14
Mann-Whitney test result for all and for each source model, both effectiveness and efficiency (U/M are artifact-based and u/m are human-perceived).

Model	UEffec (*sig*)	uEffec (*sig*)	MEffec (*sig*)	mEffec (*sig*)	UEffic (*sig*)	MEffic(*sig*)
All	0.000	0.000	0.000	0.001	0.001	0.000
M1	0.866	0.166	0.000	0.166	0.001	0.000
M2	0.000	0.031	0.009	0.502	0.550	0.110
M3	0.000	0.000	0.000	0.009	0.000	0.940
M4	0.000	0.000	0.000	0.019	0.000	0.762
M5	0.000	0.006	0.000	0.212	0.050	0.008

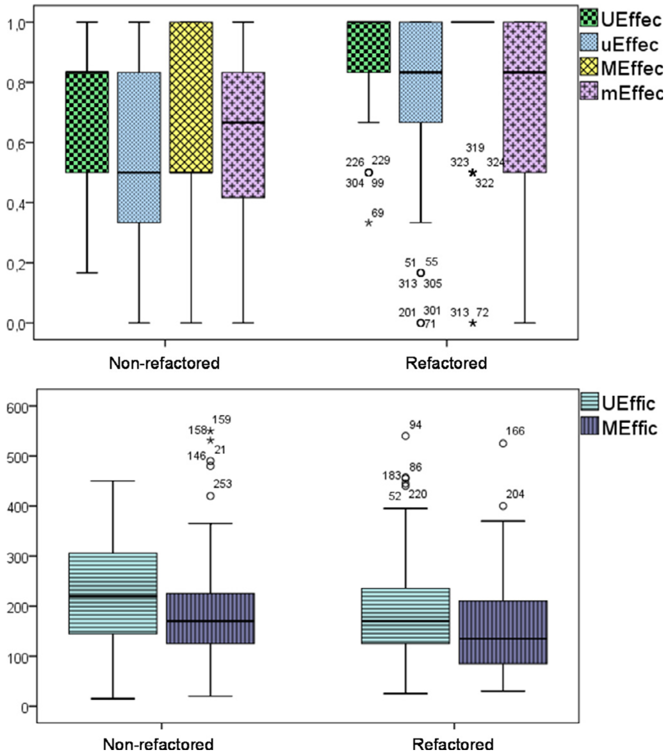


Fig. 3. Boxplots for all experimental materials (U/M are artifact-based and u/m are human-perceived).

mEffec in both treatments (with or without refactoring). Fig. 3 illustrates this assertion in diagram form, by means of a set of boxplots for the dependent variables tested.

With regard to particular business process models, the Mann-Whitney tests provide different findings. Table 14 shows these disaggregated results, in which not all the null hypotheses are rejected for some models. The null hypotheses that hold are highlighted in Table 14. The model that is most affected in this respect is M1, in which only the objective modifiability (*MEffec*) is significantly different according to the treatments.

After the comparison between Table 9 (descriptive statistics) and Table 14, it is possible to draw further conclusions. Table 9 revealed that the mean of *UEffec* and *uEffec* for M1 is fairly similar under both treatments, which is in line with the absence of a significant difference between treatments derived from the Mann-Whitney test. This is also the case for M1, in which *mEffec* was higher after refactoring. However, Table 14 reveals that the null hypothesis cannot be rejected with a 95% confidence level. The same occurs with the subjective modifiability (*mEffec*) in M2 and M5. They are statistically equal, although Table 9 shows that refactoring provides slightly better results. The remaining cases reject the null hypothesis and verify the influence of refactoring.

The second issue related to G1 concerns assessing how the efficiency varies during the understanding and modification of (refactored and non-refactored) business process models. A set of Mann-Whitney tests was similarly performed for *UEffic* and *MEffic*, following the same hypotheses.

Table 14 also shows the result of the Mann-Whitney tests by aggregating the results for all the models. The significance level is again less than 0.05 in the entire set of cases; the distributions are therefore different for each treatment. Moreover, descriptive statistics (see Table 10) demonstrated that both the understanding and the modification of refactored business process models are less time-consuming than the understanding and modification of non-refactored ones. Fig. 3 also illustrates this assertion by means of a boxplot containing the dependent and independent variables being studied.

In addition, the Mann-Whitney tests verified that there are significant differences between these two treatments throughout the five business process models used in the experiment (see Table 14). Despite the fact that the descriptive statistics in Table 10 show that the understanding/modifying time is less in the case of refactored business process models, these hypotheses contrast tests which reveal that some of those differences are not really significant to a confidence level of 95% (see highlighted cells in Table 14). For example, for model M3, the difference is not significant as regards *MEffic*, although the mean time required to modify the refactoring business process model was less than that required to modify the non-refactored one; 226 and 242 seconds, respectively (see Table 10).

5.3.2. G2. Correlation between artifact-based and human-perceived results

The second research goal, G2, is addressed in this section, which is devoted to verifying the correlation between the artifact-based understandability and modifiability values based on quantified measures obtained from literature, and the understandability and modifiability perceived by the subjects involved in this experiment.

Please recall that Table 3 showed the understandability and modifiability by means of the mean of the size, connectivity, density, separability and depth of the experimental material before and after applying refactoring operators.

The perceived value of these quality features was, on the other hand, shown in Table 9. The intensity of the linear correlation between artifact-based and perceived variables is quantified by the Spearman linear correlation test. There are two linear regression models. The first considers *UEffec* and *uEffec* vs *size, connectivity, separability, density* and *depth*. The second considers *MEffec* and m*Effec* vs c*onnectivity, density* and *separability* as the *independent variables*.

Table 15 shows the Spearman's correlation coefficient ρ (between −1 and 1) obtained for each pair of variables. ρ indicates the degree to which the values of one variable are close to the values of the other, while *sig* values lower than 0.05 mean that the pair of variables are correlated with a confidence level of 95%. The table shows that all pairs of variables are correlated. These data reveal that there is an inverse correlation between *Size, Separability* and *Depth*, and the perceived understandability and modifiability, since the value of ρ is negative, i.e., *Size, Separability* and *Depth* negatively affect understandability, while *Separability* negatively affects modifiability. These findings coincide with the assumptions made in previous works regarding *Size* and *Depth,* but there is no consensus regarding separability. Unexpectedly,

Table 15
Spearman linear correlation values for all and for each source model (U/M are artifact-based and u/m are human-perceived).

	UEffec		uEffec		MEffec		mEffec	
	ρ	*sig*	ρ	*sig*	ρ	*sig*	ρ	*sig*
Size	−0.598	0.000	−0.454	0.000	–	–	–	–
Connectivity	0.124	0.025	0.203	0.000	0.328	0.000	0.181	0.001
Density	0.473	0.000	0.404	0.000	0.501	0.000	0.252	0.000
Separability	−0.574	0.000	−0.459	0.000	−0.462	0.000	−0.262	0.000
Depth	−0.422	0.000	−0.291	0.000	–	–	–	–

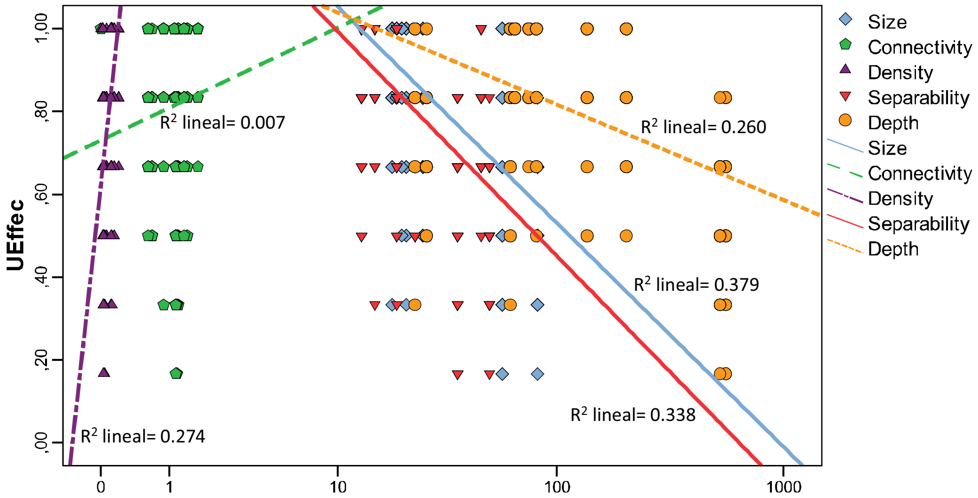


Fig. 4. Correlation between hypothesized variables and *UEffec* (artifact based).

however, *Connectivity* affects understandability and modifiability positively, i.e., business process models with higher connectivity are more understandable and modifiable. The same occurs in the case of density: it affects the understandability and modifiability of a business process model positively, i.e., business process models with greater density are more understandable and modifiable. What is more, the correlation degree (ρ) is higher for density than for connectivity.

Although the significance values (*sig*) reveal that there are correlations, these are weak in most cases. The strongest linear correlation proved to be between *UEffec* and *Size,* with a ρ that was only $= -0.598$.

We should point out that Fig. 4, Fig. 5, Fig. 6 and Fig. 7 show the scatterplot for each pair of variables with a logarithmic scale. These figures show the weak linear relation between every pair of variables in the form of a diagram.

6. Discussion

This section provides a discussion and interpretation of the findings obtained in the previous analysis and is divided into 3 subsections. The first sub-section explains the chains of evidence and their connections with the previous research goals and motivations established at the beginning of this paper. The second focuses on the threats to validity. Finally, the inferences and lessons learned are provided in the third sub-section.

6.1. Evaluation of results and implications

This subsection explains the findings of the data analysis shown in the previous section. It is in turn divided into the two subparts considered in Section 5.2 and 5.3.

6.1.1. G1. Effectiveness and efficiency results

Bearing in mind the results shown in Section 5.3.1 and obtained by means of Mann-Whitney tests, it was proven that the refactored business process models were more understandable and modifiable than the original models. It was additionally proven that there was a significant difference between managing refactored models and managing non-refactored ones. Null hypotheses $H_{0_{11}}$and $H_{0_{12}}$cannot, therefore, be rejected; in fact, they can be accepted, since refactored business process models are understood and modified more effectively than non-refactored ones. Nonetheless, these results have to be treated carefully, since some differences were observed between the business process models used in the experimental material, and a future replication might therefore be necessary.

Similarly, the results of the Mann-Whitney tests carried out in Section 5.3.1 signify that null hypotheses $H_{0_{13}}$and $H_{0_{14}}$cannot be rejected. These hypotheses are thus accepted, which implies that refactored business process models are understood and modified more efficiently than non-refactored ones.

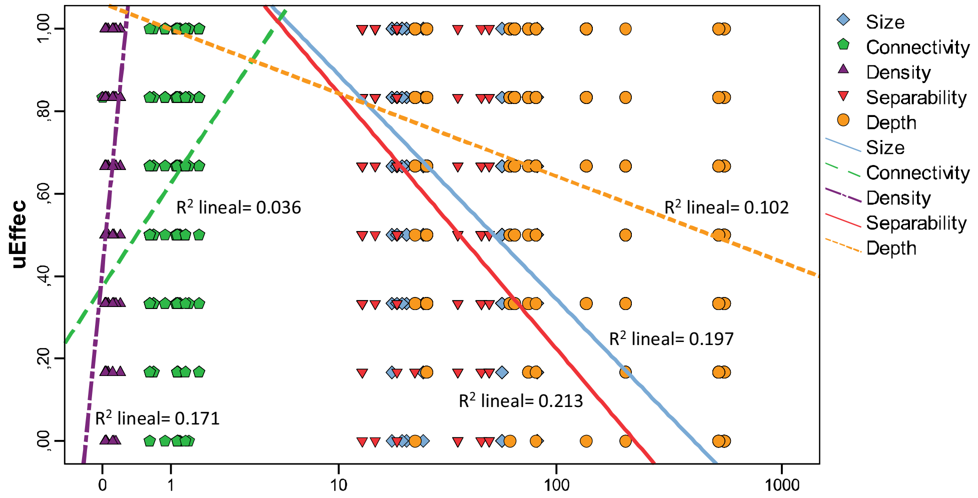


Fig. 5. Correlation between hypothesized variables and *uEffec* (human-perceived).

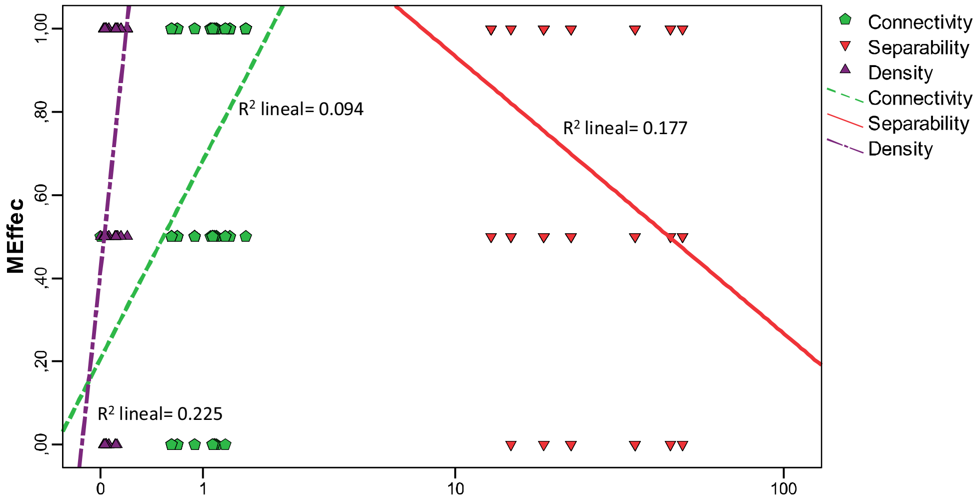


Fig. 6. Correlation between hypothesized variables and *MEffec* (artifact-based).

6.1.2. G2. Correlation between artifact-based and human-perceived results

The Spearman linear correlation tests shown in Section 5.3.2 preliminary investigates some assumptions regarding the understandability and modifiability measured by using metrics taken from the relevant literature (i.e., size, connectivity, separability, density and depth). The Spearman correlation tests points out a correlation between these measures and the artifact-based and human-perceived understandability/modifiability. However, the sign of the correlation is not coherent with the assumptions made in previous works.

Table 15 showed the correlation between the variables cited. Size negatively affects understandability. Connectivity unexpectedly had a positive effect on understandability and modifiability. Likewise, in the case of density, this measure had a positive effect as regards understanding and modifying a business process model. However, separability had a negative correlation as regards the understandability and modifiability of a business process model, a finding that goes against previous assumptions. Finally, depth was found to be a negative factor when understanding a business process model, as previous work had suggested.

The above finding simply shows that a small business process model is more understandable than a large one. It is also obvious that more connectivity between its elements makes the model easier to understand and modify than when there is less connectivity. Likewise, business process models with a high density are easier to understand and modify than models with a low density. Business process models with a high separability are more difficult to understand and modify than models with a low separability. Furthermore, business process models that are shallower are more understandable than deeper business process models.

The findings after the statistical analysis mean that the null hypothesis H_{021}cannot be rejected; artifact-based understandability and modifiability values follow the same trend as the understandability and modifiability perceived by the subjects, since they are correlated. However, connectivity and density have a positive effect as regards understanding and modifying a business process model, while separability was demonstrated to have a negative effect on these measures. This contradicts the assumptions proposed by authors in previous works. Nevertheless, although the significance values revealed that there are correlations between variables, these were weak in most cases, and it is not therefore possible to refute the assumptions made in previous work.

6.2. Threats to validity

Once the experiment had been carried out, certain issues needed to be considered as threats to its validity. There are four types of threats to validity:

- *Construct validity*: The measures chosen to quantify understandability and modifiability could have had an impact on the

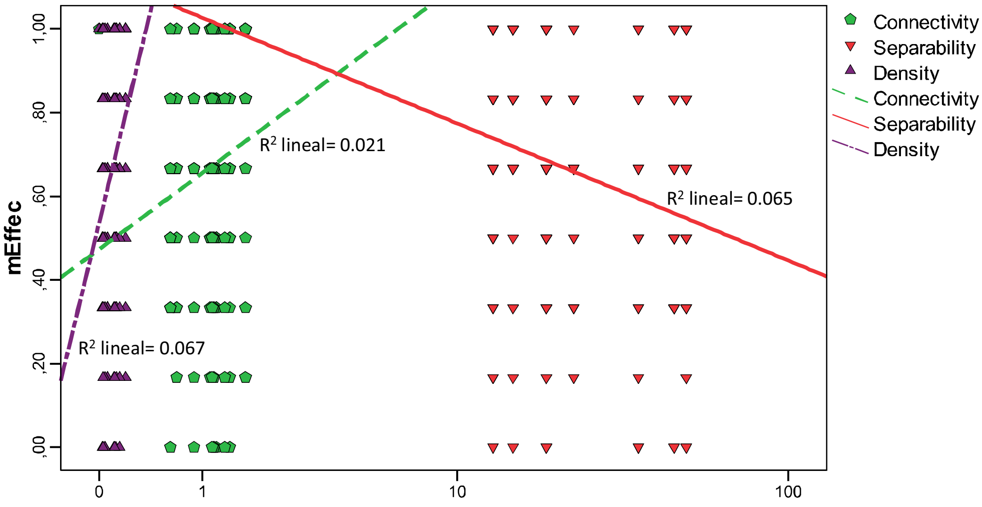


Fig. 7. Correlation between hypothesized variables and *mEffec* (human-perceived).

results of the correlation hypothesis. Although it is certainly the case that these measures are well-known and have been widely used in experiments throughout the related literature, the choice and definition of these variables may be a threat. Moreover, the positive or negative impact of these measures on understandability and modifiability could also be a threat. Hence, other statements from alternative works should be considered. Other threat to the construct validity is the lack of theoretical proof about the effect of the used measures on the modifiability and understandability of business process models. As regards the tasks defined, these were homogeneously defined in all the different business process models, and proposed with sufficient complexity to obtain significant results. Moreover, all the questionnaires in the proposed tasks follow a standard form and their responses are based on five- and seven-point scales. Finally, social threats such as evaluation apprehension have been mitigated, since the students were not graded on their performance in the experiment.

- *Internal validity*: The within-subjects balanced design of the experiment was defined with the aim of avoiding internal threats. The grouping was carried out according to the students' skills; the material was distributed in both groups following a Latin square strategy, without any advantage being given to any of the treatments being compared. Furthermore, the goals of the experiment and the statistical analysis to be performed were clearly defined at the beginning. In addition, deviations from the experiments were solved "on the fly", as mentioned above. The experimental material was based on a set of questionnaires and a few open questions. The execution of the experiment was designed in such a way as to attempt to ensure that the instructions of the experiment were well understood by the subjects, thanks to the preliminary example of it that they were given. An additional threat to the internal validity is the fatigue of the subjects during the realization of the experiment. This threat could be reduced in future replications by providing more experimental units arranged in more subject groups so that subjects have to deal with fewer tasks during the experiment.
- *External validity*: It is true that students might not properly represent the intended user population, but the tasks in the experiment were designed in such a way that no great industrial expertise was required, while simultaneously being of a sufficient complexity to obtain significant differences. However, it would be interesting to replicate the experiment with business experts and software practitioners. Another major threat concerns the experimental material and tasks defined for this material. This material was produced from two specific existing information systems by using MARBLE, a specific reverse engineering technique and its supporting tool, and IBUPROFEN was also chosen as the business process refactoring tool. The use of these techniques/tools limits the generalization of results since the original material and the refactored material are strongly dependent on both techniques. In other words, business process models could be different in case other reverse engineering tool is used on the same information systems. A further replication which would consider alternative reverse engineering and refactoring techniques must therefore be conducted in order to mitigate this threat and strengthen the results. The set of refactoring operators applied is also a threat to validity. The order of application may provide different results, as was demonstrated in M. Fernández-Ropero et al. (2013). The translation from English to Italian could be another threat that needs to be mitigated by means of, for example, the replication of the experiment using English speakers.
- *Conclusion validity*: The statistical tests were used to accept the null hypotheses. However, alternative tasks in the experiments may provide more data and more precise results. Moreover, although the correlation test revealed the linear correlation between hypothesized and perceived understandability and modifiability, the significance of such correlations was weak in all cases. It is not, therefore, possible to obtain reliable conclusions, signifying that this experiment needs to be replicated in order to compare results.

6.3. Inferences and lessons learned

Several studies had previously assessed the positive effect of refactoring on the understandability and modifiability of business process models. These evaluations were carried out by means of measures related to these quality features. Although these studies demonstrated the relationships between the size, connectivity, separability, density and depth measures and the understandability and modifiability of business process models, experiments involving people should have been performed to assess the understandability and modifiability perceived by them after refactoring. There are two main inferences of this work:

- Refactoring is a technique that improves the degree of quality of business process models, since humans find it easier to understand and modify them. This implies that business process models with a higher degree of quality provide more benefits for requirement elicitation and analysis and therefore for enterprise management and software development.
- There is a correlation between the artifact-based and human-perceived understandability and modifiability of business process models: this implies that measures used in previous artifact-based experiments have been validated once more, since they correlate with the quality perceived by humans. The most valuable finding was the unexpected positive correlation between connectivity and density with regard to understandability and modifiability and the negative relationship between separability and these features. This finding contradicts previous assumptions, but there are not sufficient data to accept this finding as true.

The post-test conducted by the subjects provided us with additional and valuable feedback that needs to be taken into account in future replications. Fig. 8 summarizes the results obtained in the post-test. There is an insight suggesting that the difficulty of the questionnaires was considered normal- neither easy nor difficult. Regarding the background provided, the subjects expressed that it was suitable for solving the questionnaires. The purpose of the questionnaires was also clear or very clear for the subjects. However, 19% of the subjects considered the experimental tasks to be unclear. The experimental material was considered unclear by only 13% of the population. It should also be noted that the time needed to perform the experiment was considered adequate by all the subjects.

After analyzing the feedback from the subjects, the lessons learned from the experiment are the following:

- The background lecture was suitable for the experiment to be performed properly.
- The subjects did not consider the questionnaires to be too difficult. This implies that in a future replication the difficulty of the tasks can be increased. However, according to the feedback from the subjects, the translation into Italian was a problem; the clarity of the tasks was thus affected. An experiment therefore needs to be carried out entirely in English to mitigate this threat.
- The material was also clear for the subjects. This suggests that larger and more intricate business process models can be used in future experiments or replications.
- The purpose of the questionnaires was well-defined, as the post-test results highlight.
- The subjects considered the questions in each task to be clear.
- According to the subjects, the time needed to perform the experiment was sufficient. This implies that it is feasible to incorporate more tasks into the questionnaires in order to obtain stronger results.

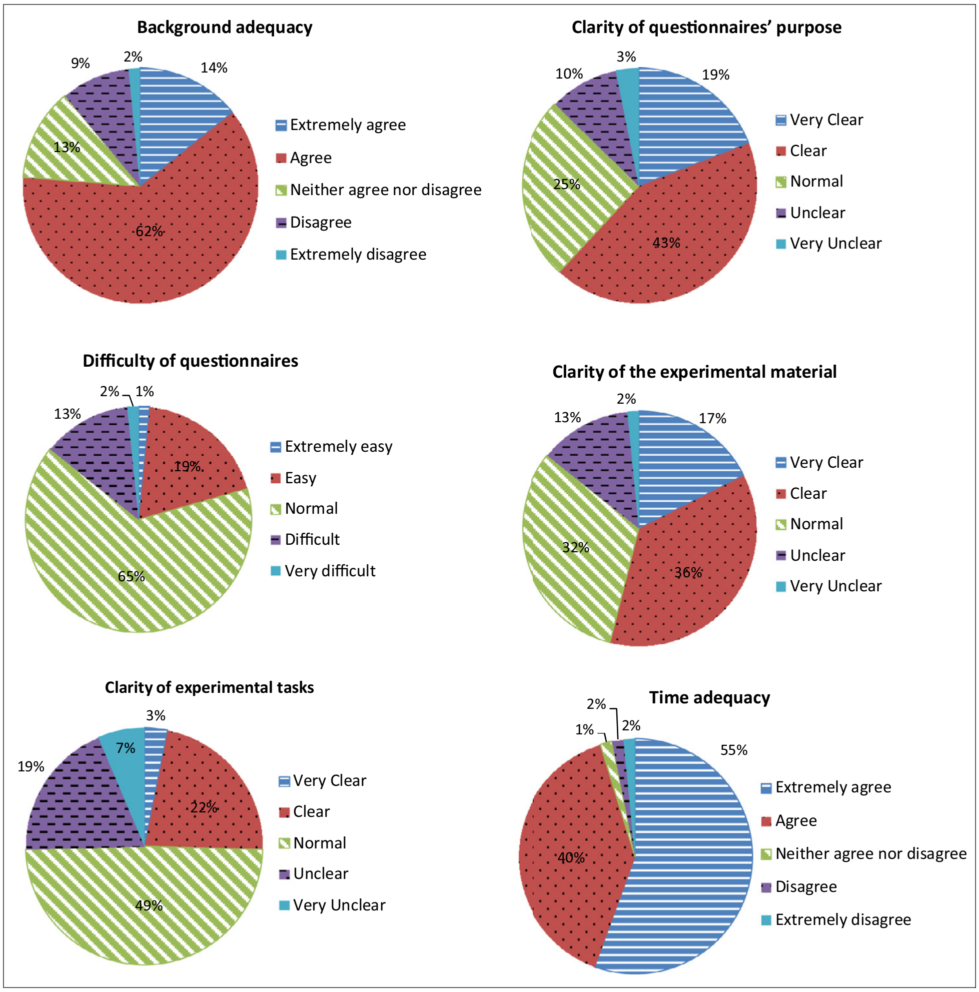


Fig. 8. Post-test results.

All this feedback will help in the future replication of the experiment and contribute to it being possible to perform other experiments in a similar research area.

6.4. Practical implications for researchers and practitioners

After conducting this experiment and analyzing results some lessons learned can be summarized as general implications for researchers and practitioners:

- Business Process models obtained from Information Systems by reverse engineering can be improved and sanitized by means of refactoring. Despite of semantic lost and other drawbacks associated with reverse engineering techniques, business process models can be obtained using such techniques since it can be complemented with refactoring by executing both in a row.
- Both analyzed quality characteristics (understandability and modifiability) are directly related to the effort necessary to maintain business process models. Having refactored business process models can help to save cost regarding the maintainability of such models.
- Having analyzed the correlation between artifact-based and human-perceived measures, business process modelling tools could implement and use some indicators using artifacts-based measures (i.e., size, connectivity, density, separability) to estimate the understandability and modifiability levels concerning human viewpoint.

7. Conclusions and future work

Refactoring techniques have been used in literature to increase the degree of quality of business process models. However, none of this work has checked how refactoring affects the quality as perceived by humans, in an attempt to assess the applicability and feasibility of refactoring techniques. This work therefore presents a controlled experiment involving 65 computer science students, the purpose of which was to assess the influence of refactoring on the perceived quality of business process models.

Understandability and modifiability were evaluated by means of several tasks with several experimental materials obtained from real-world information systems. This material was analyzed by subjects with and without applying refactoring operators, and the results were then used in statistical tests. The aim of these statistical tests was to discover the effect of refactoring on business process models. The findings obtained by means of these statistical tests were that refactored business process models are more effectively and efficiently understood and modified than non-refactored business process models, i.e., refactoring reduces the time needed to perform actions, and its results are of a higher quality. Moreover, the experiment gave extra strength to the assertion that hypothesized variables of understandability and modifiability are correlated with perceived understandability and modifiability.

Several conclusions may be inferred after the conducted experiment. Since business process models proved to be more understandable and modifiable according to human perception, the main implication is that refactoring is appropriate for dealing with ordinary quality faults that prevent the understandability and modifiability of business process models as obtained by means of reverse engineering. Moreover, some assumptions concerning measures with which to assess understandability and modifiability were verified. Small business process models were proven to be more understandable than large ones. However, we observed that more connectivity between elements of business process models makes them easier to understand and modify than less connectivity; this contrasts with previous assumptions. Similarly, business process models with a high density proved to be easier to understand and modify than models with a small density. Business process models with a low separability were seen to be easier to understand and modify than models with a high separability. Moreover, it is clear that business process models that were shallower were more understandable than deeper business process models.

The unexpected positive relation between connectivity and density with regard to understandability and modifiability contradicted the assumptions made in previous work, as did the negative relationship between separability and these quality features. However, there are insufficient data to ensure that this finding is true, and more experimentation is required to shed light on this (Cardoso, 2006).

The main implication derived from the conclusions extracted from the analysis of G1 is that business process refactoring should be used to achieve more understandable and modifiable business process models, especially when these models are obtained by means of reverse engineering and therefore have recurrent quality faults. Additionally, the conclusions drawn from G2 lead to a second major implication, pointing to the fact that the understandability and modifiability of business process models cannot be assessed by trusting only in intrinsic, quantifiable measures. Accurate values of understandability and modifiability can instead be achieved by combining those measures with human-perceived assessment. Furthermore these results should be kept into consideration in practice during job rotation in software organizations that adopt business process refactoring (Santos et al., 2017).

Taking into account the threats to validity mentioned, our future research work will consist of the replication of the experiment, using experts as experimental units in order to generalize the results. Other variables and supporting tools may also be considered, along with additional hypotheses concerning refactoring.

Acknowledgments

This work is part of the project GINSENG(TIN2015-70259-C2-1-R) funded by the Spanish Ministry of Economy and Competitiveness and the FEDER Fund (*Fondo Europeo de Desarrollo Regional*); project SOS (SBPLY/17/180501/000364) funded by *"Consejería de Educación, Cultura y Deportes de la Dirección General de Universidades, Investigación e Innovación de la JCCM";* and also supported by the University of Bari and the grant that Dr. Ricardo Pérez del Castillo enjoys from JCCM within the initiatives for talent retention and return in line with RIS3 goals.

Appendix. I

Table 16 shows each measure and its association with the characteristics of understandability and modifiability, as introduced in Section 2.

Table 16
Measures of quality features.

Measure	Understandability	Modifiability	Proposed by
Total number of sequence flows	•		Rolon et al. (2009)
Total number of events	•		Rolon et al. (2009)
Total number of gateways	•		Rolon et al. (2009)
Number of sequence flows from events	•		Rolon et al. (2009)
Number of association flows	•		Rolon et al. (2009)
Number of sequence flows from gateways	•	•	Rolon et al. (2009)
Connectivity level between pools	•		Rolon et al. (2009)
Number of data objects which are outputs of activities	•		Rolon et al. (2009)
Number of data objects which are inputs of activities	•		Rolon et al. (2009)
Connectivity level between activities		•	Rolon et al. (2009)
Control flow complexity	•	•	Dumas et al. (2011), Cardoso (2006)
Size (Number of Nodes)	•		L. Sánchez-González et al. (2010), Mendling et al. (2007)
Diameter	•		L. Sánchez-González et al. (2010), Mendling et al. (2007)
Density	•	•	L. Sánchez-González et al. (2010), Mendling et al. (2007)
Connectivity	•	•	L. Sánchez-González et al. (2010), Mendling et al. (2007)
Average Gateway Degree	•		L. Sánchez-González et al. (2010)
Maximum Gateway Degree		•	L. Sánchez-González et al. (2010), Mendling et al. (2007)
Separability	•	•	L. Sánchez-González et al. (2010), Mendling et al. (2007)
Sequentiality	•	•	L. Sánchez-González et al., 2010, Mendling et al. (2007)
Depth	•		L. Sánchez-González et al. (2010), Mendling et al. (2007)
Gateway Mismatch	•	•	L. Sánchez-González et al. (2010)
Gateway Heterogeneity	•	•	L. Sánchez-González et al. (2010)

Appendix. II

This appendix shows the training example provided to the participants before the experiment was conducted. Fig. 9 shows the material for the example. This model has been obtained from the *Tabula* information system and no refactoring operator has been applied. Fig. 10 shows the questionnaires for the aforementioned material. The first questionnaire corresponds to the understandability part, while the second corresponds to the modifiability part. Each questionnaire is on a separate page.

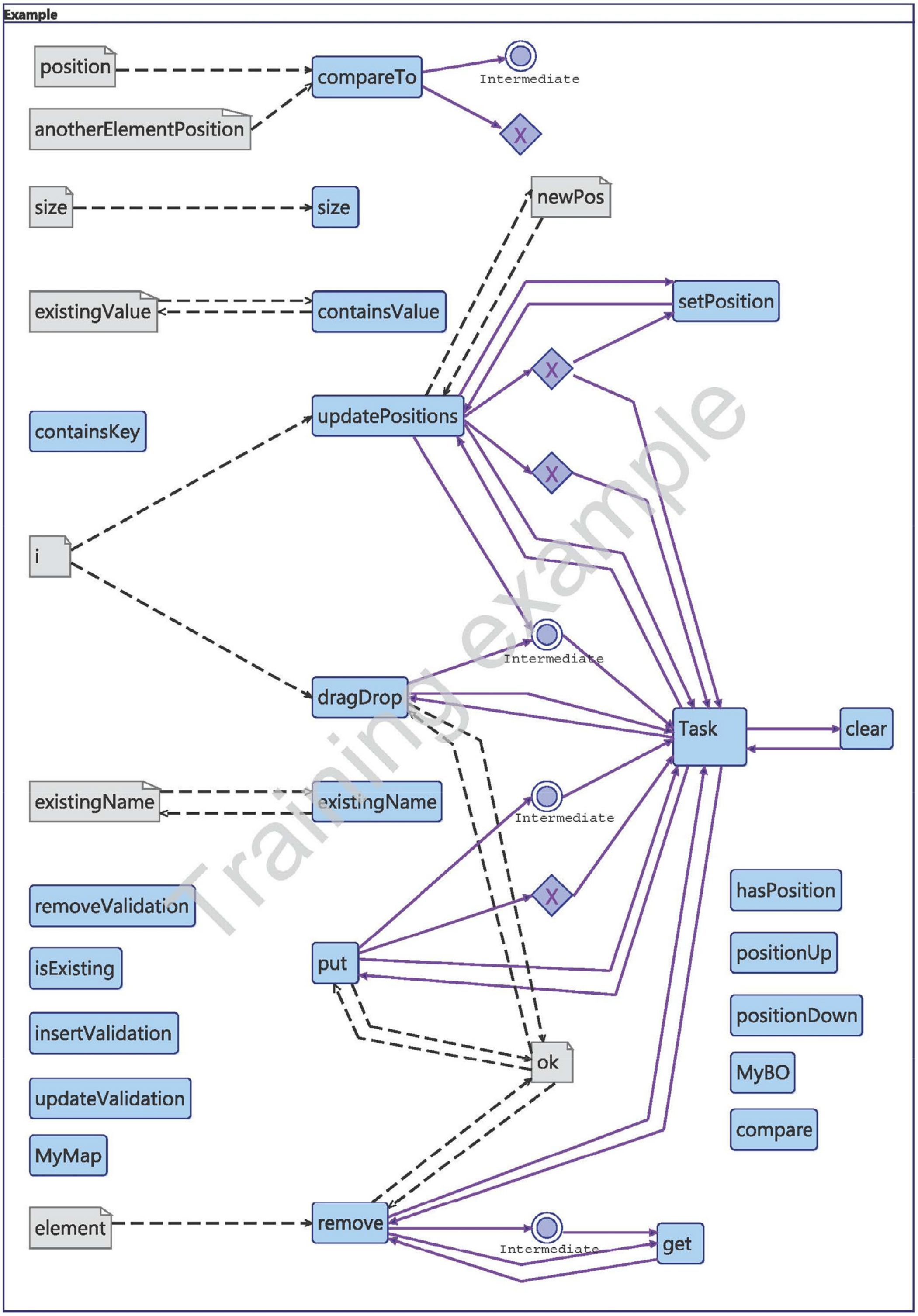
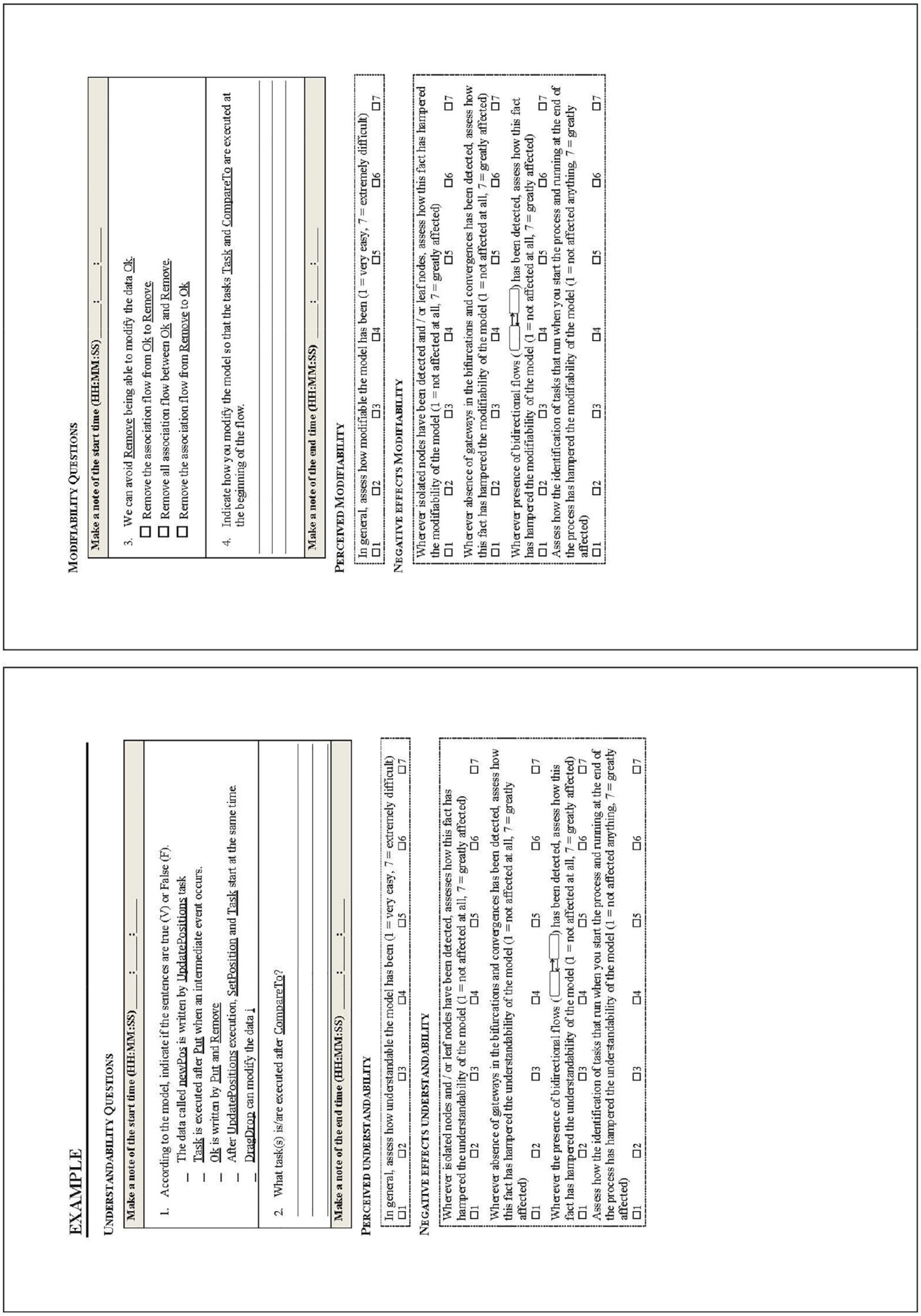


Fig. 9. Material for the training example.



EXAMPLE

UNDERSTANDABILITY QUESTIONS

Make a note of the start time (HH:MM:SS) ____ : ____ : ____

1. According to the model, indicate if the sentences are true (V) or False (F).
   - The data called newPos is written by UpdatePositions task
   - Task is executed after Put when an intermediate event occurs.
   - Ok is written by Put and Remove
   - After UpdatePositions execution, SetPosition and Task start at the same time.
   - DragDrop can modify the data i

2. What task(s) is/are executed after CompareTo?

Make a note of the end time (HH:MM:SS) ____ : ____ : ____

PERCEIVED UNDERSTANDABILITY

In general, assess how understandable the model has been (1 = very easy, 7 = extremely difficult)
□1 □2 □3 □4 □5 □6 □7

NEGATIVE EFFECTS UNDERSTANDABILITY

Wherever isolated nodes and / or leaf nodes have been detected, asseses how this fact has hampered the understandability of the model (1 = not affected at all, 7 = greatly affected)
□1 □2 □3 □4 □5 □6 □7

Wherever absence of gateways in the bifurcations and convergences has been detected, assess how this fact has hampered the understandability of the model (1 = not affected at all, 7 = greatly affected)
□1 □2 □3 □4 □5 □6 □7

Wherever the presence of bidirectional flows (⇄) has been detected, assess how this fact has hampered the understandability of the model (1 = not affected at all, 7 = greatly affected)
□1 □2 □3 □4 □5 □6 □7

Assess how the identification of tasks that run when you start the process and running at the end of the process has hampered the understandability of the model (1 = not affected anything, 7 = greatly affected)
□1 □2 □3 □4 □5 □6 □7

MODIFIABILITY QUESTIONS

Make a note of the start time (HH:MM:SS) ____ : ____ : ____

3. We can avoid Remove being able to modify the data Ok.
   - ☐ Remove the association flow from Ok to Remove.
   - ☐ Remove all association flow between Ok and Remove.
   - ☐ Remove the association flow from Remove to Ok

4. Indicate how you modify the model so that the tasks Task and CompareTo are executed at the beginning of the flow.

Make a note of the end time (HH:MM:SS) ____ : ____ : ____

PERCEIVED MODIFIABILITY

In general, assess how modifiable the model has been (1 = very easy, 7 = extremely difficult)
□1 □2 □3 □4 □5 □6 □7

NEGATIVE EFFECTS MODIFIABILITY

Wherever isolated nodes have been detected and / or leaf nodes, assess how this fact has hampered the modifiability of the model (1 = not affected at all, 7 = greatly affected)
□1 □2 □3 □4 □5 □6 □7

Wherever absence of gateways in the bifurcations and convergences has been detected, assess how this fact has hampered the modifiability of the model (1 = not affected at all, 7 = greatly affected)
□1 □2 □3 □4 □5 □6 □7

Wherever presence of bidirectional flows (⇄) has been detected, assess how this fact has hampered the modifiability of the model (1 = not affected at all, 7 = greatly affected)
□1 □2 □3 □4 □5 □6 □7

Assess how the identification of tasks that run when you start the process and running at the end of the process has hampered the modifiability of the model (1 = not affected anything, 7 = greatly affected)
□1 □2 □3 □4 □5 □6 □7

Fig. 10. Questionnaires for the training example.

Appendix. III

A Shapiro-Wilk test was used to verify the normality of the *UEffec, uEffec, MEffec, mEffec, UEffic* and *MEffic* across levels of *T*. Hypotheses for this test are as follows:

H_0: The sample follows a normal distribution
H_1: The sample does not follow a normal distribution

If the statistical significance (*p-value*) is lower than 0.05, the null hypothesis is therefore rejected and the alternative one is accepted. Table 17

Table 17
Shapiro-Wilk test overall.

T	UEffec		uEffec		MEffec		mEffec		UEffic		MEffic	
	p-value	*sig*	*p-value*	*sig*	*p-value*	*sig*	*p-value*	*sig*	*p-value*	*sig*	*p-value*	*sig*
0	0.9052	0.0000	0.9350	0.0000	0.8050	0.0000	0.9258	0.0000	0.9171	0.0000	0.8918	0.0000
R	0.6978	0.0000	0.8669	0.0000	0.4990	0.0000	0.8751	0.0000	0.9148	0.0000	0.9223	0.0000

Table 18
Shapiro-Wilk test for each source model.

Model	T	UEffec		uEffec		MEffec		mEffec		UEffic		MEffic	
		p-value	*sig*	*p-value*	*sig*	*p-value*	*sig*	*p-value*	*sig*	*p-value*	*sig*	*p-value*	*sig*
M1	0	0.6012	0.0000	0.7524	0.0000	0.6889	0.0000	0.8137	0.0000	0.9534	**0.1549**	0.8133	0.0000
	R	0.6399	0.0000	0.8218	0.0001	0.2698	0.0000	0.7232	0.0000	0.7763	0.0000	0.8701	0.0014
M2	0	0.8152	0.0000	0.9215	0.0258	0.8033	0.0001	0.9070	0.0108	0.9548	**0.2118**	0.9732	**0.6108**
	R	0.7550	0.0000	0.9171	0.0134	0.6577	0.0000	0.9223	0.0188	0.9083	0.0088	0.8593	0.0005
M3	0	0.8238	0.0000	0.9537	**0.1589**	0.7721	0.0000	0.9361	0.0471	0.9644	**0.3249**	0.7870	0.0000
	R	0.8238	0.0000	0.8565	0.0007	0.2698	0.0000	0.9178	0.0207	0.9121	0.0146	0.8513	0.0005
M4	0	0.8989	0.0068	0.9415	**0.0911**	0.7623	0.0000	0.9338	**0.0558**	0.8481	0.0005	0.8977	0.0063
	R	0.7167	0.0000	0.7266	0.0000	0.3778	0.0000	0.8679	0.0007	0.8980	0.0041	0.9457	**0.0913**
M5	0	0.8930	0.0030	0.9193	0.0175	0.7453	0.0000	0.9229	0.0249	0.9386	**0.0559**	0.9135	0.0107
	R	0.7957	0.0000	0.9096	0.0145	0.6614	0.0000	0.8395	0.0003	0.9626	**0.3420**	0.9547	**0.2107**

shows the value of the statistical significance. The *p-value* was lower than 0.05 when data were studied all together without division by experimental material. This means that the null hypothesis can be rejected and the distribution is not normal, with a significance level of 95%. However, when the division by experimental material is achieved, the *p-value* is greater than 0.05 in some cases (see bold cells in Table 18), which means that the null hypothesis cannot be rejected in these cases. It was not possible to apply parametric tests in this experiment.

References

Abrahão, S., Insfran, E., Carsí, J.A., Genero, M., 2011. Evaluating requirements modeling methods based on user perceptions: a family of experiments. Inf. Sci. 181 (16), 3356–3378.

Baldassarre, M.T., Carver, J., Dieste, O., Juristo, N., 2014. Replication types: towards a shared taxonomy. In: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. ACM.

Basili, V.R., Caldiera, G., Rombach, H.D., 1994. The goal question metric approach. Encycl. Softw. Eng. 2 (1994), 528–532.

Bianchi, A., Caivano, D., Visaggio, G., 2000. Method and process for iterative re-engineering of data in a legacy system. In: Seventh Working Conference on Reverse Engineering, 2000. Proceedings. IEEE.

Caivano, D., 2005. Continuous software process improvement through statistical process control. In: Ninth European Conference on Software Maintenance and Reengineering, 2005. CSMR 2005. IEEE.

Caivano, D., Lanubile, F., Visaggio, G., 2001. Software renewal process comprehension using dynamic effort estimation. In: IEEE International Conference on Software Maintenance, 2001. Proceedings. IEEE.

Canfora, G., Penta, M.D., Cerulo, L., 2011. Achievements and challenges in software reverse engineering. Commun. ACM 54 (4), 142–151. http://dx.doi.org/10.1145/1924421.1924451.

Cardoso, J., 2006. Process control-flow complexity metric: an empirical validation. In: IEEE International Conference on Services Computing. Chicago, IL, 8-22 Sept. 2006. pp. 167–173. http://dx.doi.org/10.1109/SCC.2006.82.

Carver, J.C., Juristo, N., Baldassarre, M.T., Vegas, S., 2013. Replications of software engineering experiments. Empirical Software Eng. 19 (2), 267–276. http://dx.doi.org/10.1007/s10664-013-9290-8.

Conforti, R., Dumas, M., García-Bañuelos, L., La Rosa, M., 2014. Beyond tasks and gateways: discovering BPMN models with subprocesses, boundary events and activity markers. Business Process Management. Springer, pp. 101–117.

Di Francescomarino, C., Marchetto, A., Tonella, P., 2009. Reverse engineering of business processes exposed as web applications. In: 13th European Conference on Software Maintenance and Reengineering (CSMR'09). Kaiserslautern, Germany. IEEE Computer Society: Fraunhofer IESE, pp. 139–148.

Dijkman, R., Gfeller, B., Küster, J., Völzer, H., 2011. Identifying refactoring opportunities in process model repositories. Inf. Software Technol.

Dijkman, R., Rosa, M.L., Reijers, H.A., 2012. Managing large collections of business process models—current techniques and challenges. Comput. Ind. 63 (2), 91.

Dumas, M., García-Bañuelos, L., Rosa, M.L., Uba, R., 2011. Clone detection in repositories of business process models. In: Proceedings of the 9th International Conference on Business Process Management. Clermont-Ferrand, France. Springer-Verlag, pp. 248–264.

Ekanayake, C.C., Dumas, M., García-Bañuelos, L., La Rosa, M., ter Hofstede, A.H., 2012. Approximate clone detection in repositories of business process models. Business Process Management. Springer, pp. 302–318.

Fernández-Ropero, M., Pérez-Castillo, R., Boffoli, N., Caivano, D., Piattini, M., 2013. *Extra material of Empirical Study on Understandability and Modifiability of Refactored Business Process* Available from: http://alarcos.esi.uclm.es/per/mfernandez/material5.html.

Fernández-Ropero, M., Pérez-Castillo, R., Cruz-Lemus, J.A., Piattini, M., 2013b. Assessing the best-order for business process model refactoring. In: Proceedings of the 28th Annual ACM Symposium on Applied Computing. Coimbra, Portugal, March 18 - 22. pp. 1397–1402. http://dx.doi.org/10.1145/2480362.2480625.

Fernández-Ropero, M., Pérez-Castillo, R., Cruz-Lemus, J.A., Piattini, M., 2013c. Assessing the best-order for business process model refactoring. In: 28th Symposium On Applied Computing (SAC). Coimbra, Portugal. pp. 1400–1406.

Fernández-Ropero, M., Pérez-Castillo, R., Piattini, M., 2013d. Challenges of business process model improvement after reverse engineering. In: SEM 2013 in Conjunction with ENASE. Angers, France.

Ferreira, W., Baldassarre, M.T., Soares, S.Codex, 2018. A metamodel ontology to guide the execution of coding experiments. Comput. Stand. Interfaces 59, 35–44. http://dx.doi.org/10.1016/j.csi.2018.02.003.

Gambini, M., La Rosa, M., Migliorini, S., Ter Hofstede, A., 2011. Automated error correction of business process models. Bus. Process Manage. 148–165.

Indulska, M., Recker, J., Rosemann, M., Green, P., 2009. Business process modeling: current issues and future challenges. Advanced Information Systems Engineering. Springer.

ISO/IEC, 2011. ISO/IEC 25010:2011. In: Systems and software engineering – System and software product Quality Requirements and Evaluation (SQuaRE) – System and software quality models.

Jedlitschka, A., Ciolkowski, M., Pfahl, D., 2008. Reporting experiments in software engineering. Guide to Advanced Empirical Software Engineering. pp. 201–228.

Juristo, N., Moreno, A.M., 2013. Basics of Software Engineering Experimentation. Springer Science & Business Media.

La Rosa, M., Wohed, P., Mendling, J., ter Hofstede, A.H.M., Reijers, H.A., van der Aalst, W., 2011. Managing process model complexity via abstract syntax modifications. Ind. Inf. IEEE Trans. 7 (4), 614–629.

Leopold, H., Smirnov, S., Mendling, J., 2010. Refactoring of process model activity labels. In: Proceedings of the Natural language processing and information systems, and 15th international conference on Applications of natural language to information systems. Cardiff, UK. Springer-Verlag, pp. 268–276.

Leopold, H., Smirnov, S., Mendling, J., 2012. On the refactoring of activity labels in business process models. Inf. Syst. 37 (5), 443–459.

Marshall, M.N., 1996. Sampling for qualitative research. Fam. Pract. 13 (6), 522–525. https://doi.org/10.1093/fampra/13.6.522.

Mendling, J., Reijers, H., Cardoso, J., 2007. What makes process models understandable? Bus. Process Manage. 48–63.

Mendling, J., Strembeck, M., 2008. Influence Factors of Understanding Business Process Models. Springer.

Normantas, K., Vasilecas, O., 2013. A systematic review of methods for business knowledge extraction from existing software systems. Baltic J. Mod. Comput. 1 (1) p.

InPress.

Nugroho, A., 2009. Level of detail in UML models and its impact on model comprehension: a controlled experiment. Inf. Software Technol. 51 (12), 1670–1685.

OMG. 2011. *Business Process Modeling Notation Specification 2.0* Available from: http://www.omg.org/spec/BPMN/2.0/PDF/.

Oppenheim, A.N., 2000. Questionnaire Design, Interviewing and Attitude Measurement. Continuum International Publishing Group ISBN: 9780826451767.

Pérez-Castillo, R., Caivano, D., Piattini, M., 2014. Ontology-based similarity applied to business process clustering. J. Software: Evol. Process 26 (12), 1128–1149. http://dx.doi.org/10.1002/smr.1652.

Pérez-Castillo, R., de Guzmán, I.G.-R., Piattini, M., 2011a. Business process archeology using MARBLE. Inf. Software Technol. 53 (10), 1023–1044. http://dx.doi.org/10.1016/j.infsof.2011.05.006.

Pérez-Castillo, R., Fernández-Ropero, M., Guzmán, I.G.-R.d., Piattini, M., 2011b. MARBLE. A business process archeology tool. In: 27th IEEE International Conference on Software Maintenance (ICSM 2011). Williamsburg, VI. pp. 578–581. http://dx.doi.org/10.1109/ICSM.2011.6080834.

Pérez-Castillo, R., Fernández-Ropero, M., Piattini, M., Caivano, D., 2013. How does refactoring affect understandability of business process models? In: 25th International Conference on Software Engineering & Knowledge Engineering (SEKE'13). Knowledge Systems Institute Graduate School: Boston, USA, pp. 644–649 DOI: ISBN-13: 978-1-891706-33-2.

Pittke, F., Leopold, H., Mendling, J., 2013. Spotting terminology deficiencies in process model repositories. Enterprise, Business-Process and Information Systems Modeling. Springer, pp. 292–307.

Polyvyanyy, A., Smirnov, S., Weske, M., 2010. Business process model abstraction. Handbook On Business Process Management 1. pp. 149–166.

Reijers, H.A., Mendling, J., 2011. A study into the factors that influence the understandability of business process models. IEEE Trans. Syst. Man Cybern. - Part A: Syst. Hum. 41 (99), 1–14.

Ricca, F., Di Penta, M., Torchiano, M., Tonella, P., Ceccato, M., 2007. The role of experience and ability in comprehension tasks supported by UML stereotypes. In: ICSE. Citeseer.

Rolon, E., Sánchez-González, L., Garcia, F., Ruiz, F., Piattini, M., Caivano, D., Visaggio, G., 2009. Prediction Models for BPMN Usability and Maintainability. IEEE.

Sánchez-González, L., García, F., Ruiz, F., Piattini, M., 2013. Toward a quality framework for business process models. Int. J. Coop. Inf. Syst. 22 (01).

Sánchez-González, L., Rubio, F.G., González, F.R., Velthuis, M.P., 2010a. Measurement in business processes: a systematic review. Bus. Process Manage. J. 16 (1), 114–134.

Sánchez-González, L., Rubio, F.G., Mendling, J., González, F.R., 2010b. Quality assessment of business process models based on thresholds. On the Move to Meaningful Internet Systems: OTM 2010. pp. 78–95.

Santos, R.E.S., da Silva, F.Q.B., Baldassarre, M.T., de Magalhães, C.V.C., 2017. Benefits and limitations of project-to-project job rotation in software organizations: a synthesis of evidence. Inf. Software Technol. 89, 78–96. http://dx.doi.org/10.1016/j.infsof.2017.04.006.

Sjøberg, D.I., Hannay, J.E., Hansen, O., Kampenes, V.B., Karahasanovic, A., Liborg, N.-K., Rekdal, A.C., 2005. A survey of controlled experiments in software engineering. IEEE Trans. Software Eng. 31 (9), 733–753.

Smirnov, S., 2012. Business Process Model Abstraction. Universitätsbibliothek.

Smirnov, S., Reijers, H., Weske, M., 2011. A Semantic Approach For Business Process Model Abstraction. Springer.

Smirnov, S., Weidlich, M., Mendling, J., 2012. Business process model abstraction based on synthesis from well-structured behavioral profiles. Int. J. Coop. Inf. Syst. 21 (01), 55–83.

van der Aalst, W.M.P., 2012. Process mining: overview and opportunities. ACM Trans. Manage. Inf. Syst. (TMIS) 3 (2), 7.

Vegas, S., Apa, C., Juristo, N., 2016. Crossover designs in software engineering experiments: benefits and perils. IEEE Trans. Software Eng. 42 (2), 120–135. http://dx.doi.org/10.1109/tse.2015.2467378.

Weber, B., Reichert, M., Mendling, J., Reijers, H.A., 2011. Refactoring large process model repositories. Comput. Ind. 62 (5), 467–486. http://dx.doi.org/10.1016/j.compind.2010.12.012.

Weske, M., 2007. Business Process Management: Concepts, Languages, Architectures. Springer-Verlag Berlin Heidelberg, Leipzig, Germany, pp. 368.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A., 2012. Experimentation in Software Engineering. Springer Science & Business Media.

Zou, Y., Hung, M., 2006. An approach for extracting workflows from e-commerce applications. In: Proceedings of the Fourteenth International Conference on Program Comprehension. IEEE Computer Society, pp. 127–136.

Zugal, S., Pinggera, J., Weber, B., Mendling, J., Reijers, H.A., 2012. Assessing the impact of hierarchy on model understandability–a cognitive perspective. Models in Software Engineering. Springer, pp. 123–133.

Danilo Caivano is graduated at the University of Bari Aldo Moro, where he also obtained his Ph.D. in 2002 and is currently assistant professor. He carries out his research in the Software Engineering Laboratory at the Department of Informatics. His research and teaching activities focus on topics related to Software Engineering with emphasis on Project and Process Management in collocated and distributed contexts and on software development, maintenance and testing. Since 2007 he is Chief Executive Officer of SER&Practices (www.serandp.com), a Spin Off company of the University of Bari that he has contributed to start up. He is actively involved in the Project Management Institute - Southern Italy Chapter (www.pmi.org) and in the International Software Engineering Research Network (isern.iese.de).

María Fernández-Ropero holds the Ph.D. degree in Computer Science from the University of Castilla-La Mancha. She works as Software Engineering at Indra Software Labs. Her research interests include architecture-driven modernization, model-driven development and business process mining. Her email is msfernandezr@indra.es

Ricardo Pérez-Castillo holds the Ph.D. degree in Computer Science from the University of Castilla-La Mancha (Spain). He works at the Information Systems and Technologies Institute at University of Castilla-La Mancha. His research interests include architecture-driven modernization, model-driven development, business process archaeology and enterprise architecture. His email is Ricardo.pdelcastillo@uclm.es

Mario Piattini is a full professor of computer science at the University of Castilla-La Mancha, Spain. His research interests include Global Software Development and Green Software. Piattini received a Ph.D. in Computer Science from the Universidad Politécnica de Madrid. His email is Mario.Piattini@uclm.es

Michele Scalera is graduated in Information Sciences at the University of Bari. He is assistant professor in Computer Science in the same university where he currently teaches courses on Computer Networks, Software Systems Integration and Test and Informatics. He is a member of the Scientific Technical Committee of Informatics Service Center and Delegated to ICT of the Jonian Department at the University of Bari. He currently is a reviewer of several journals and international congresses. His research fields are: e-learning, network security, business intelligence and knowledge discovery.