Comparing Business Value Modeling Methods: A Family of Experiments

Eric Souza[a], Ana Moreira[a], João Araújo[a], Silvia Abrahão[b], Emilio Insfran[b], Denis Silva da Silveira[c]

[a]*NOVA LINCS, Departamento de Informática, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa, PORTUGAL*
[b]*Department of Computer Systems and Computation, Universitat Politècnica de València, SPAIN*
[c]*Department of Administrative Sciences, Universidade Federal de Pernambuco, BRAZIL*

Abstract

CONTEXT: A value model is used to describe how an organization creates, delivers, and captures its business value. Value-driven development methods use the notion of "economic value exchange" to define more efficient business strategies and align Information Systems (IS) with organizational goals. Current value-driven methods are complex and there is insufficient empirical evidence regarding which of the existing methods are more effective under what circumstances. OBJECTIVE: This paper compares two different value-driven methods to provide empirical evidence regarding both their efficacy when modeling business value and their likelihood of acceptance in practice. METHOD: This goal was addressed by performing a family of three controlled experiments with a group of novice software engineers and business analysts to compare the Dynamic Value Description (DVD) method with the e3value method, with respect to their effectiveness, efficiency, perceived ease of use, perceived usefulness and intention to use. The experiment was initially performed in Spain and then replicated in Portugal and Brazil with other participants with different backgrounds. A meta-analysis was performed to aggregate the empirical findings obtained in each experiment. RESULTS: The results indicate that the DVD method is superior with respect to all the variables analyzed. CONCLUSION: The DVD method is a promising and alternative method to specify business value when compared to the well-known e3value method for the analyzed variables.

Keywords: value model, value-driven, controlled experiment

1. Introduction

Many of today's software engineering practices and research occur in a neutral value setting, where business goals, requirements, objects, tests, and defects are equally important [1]. However, most studies about the success factors of a software development project indicate that the key critical factors lie in the organizational values of the domain [1, 2]. For example, the Standish Group's CHAOS reports state that most software design defects are caused by value-oriented weaknesses [2]. Recently, the Value-Based Software Engineering (VBSE) research area has emerged, taking the concept of *value* at the forefront of software engineering decisions [3]. For example, value-based requirements engineering includes principles and practices for identifying stakeholders and the value exchange among them for the success of a system. This is achieved by modeling the *business values* and use them to guide development of the system [3, 4, 5]. Therefore, when developing information systems, for example, it is necessary that the business values are also reflected in the system.

For these reasons, an increasing number of approaches propose value models to specify business values [6]. A *value model* describes how an organization creates, delivers, and captures business value [7, 8]. Value-oriented development methods use the notion of *economic value exchange* to define more efficient business strategies and align information systems with organizational goals [9]. VBSE is still a recent research area, and so some current value-oriented methods are still immature [3], complex [10], and there is insufficient empirical evidence on which of the existing methods are most effective under what circumstances [1]. In previous work [11], we identified the need for further empirical evidence about the circumstances under which the various value-driven methods are more effective.

Email addresses: er.souza@campus.fct.unl.pt (Eric Souza), amm@fct.unl.pt (Ana Moreira), joao.araujo@fct.unl.pt (João Araújo), sabrahao@dsic.upv.es (Silvia Abrahão), einsfran@dsic.upv.es (Emilio Insfran), dsilveira@ufpe.br (Denis Silva da Silveira)

The purpose of this paper is, therefore, to report on a family of controlled experiments carried out to compare two business value modeling methods: e3value [7], which is a widely established and applied business model representation [8, 10], and the Dynamic Value Description method (DVD), with respect to their effectiveness, efficiency, perceived ease of use, perceived usefulness and intention to use.

The first controlled experiment was conducted in Spain with MSc students at Universitat Politècnica de València (UPV) [12]. The results favored the DVD method, with the exception of the perceived usefulness of the methods, for which no significant difference could be found. This was the major motivation for the exact replication subsequently performed in Portugal with MSc students at Universidade NOVA de Lisboa (UNL). In this case, all the variables evaluated favored the DVD method. The second (exact) replication took place in Brazil and aimed at validating the results from the two previous experiments with Business Management PhD students (who are also practitioners in industry) at Universidade Federal de Pernambuco (UFPE). The results of each experiment are analyzed individually and the empirical findings obtained in each experiment are aggregated by means of a meta-analysis. The results show that the efficiency and effectiveness of the DVD method are significantly higher than the current efficacy of e3value, and that the perceived ease of use, perceived usefulness and intention to use reported by the participants are also significantly higher for the DVD method.

The remainder of this paper is organized as follows. Section 2 introduces value-based development. Section 3 summarizes the two chosen methods (e3value and DVD) for our the experiments, conceptually compares them, and finishes discussing existing studies comparing value-driven methods. Section 4 presents an overview of the method used to perform the family of experiments. Section 5 presents the design of the baseline experiment performed in Spain (and later replicated in Portugal and Brazil), as well as a summary of the results with the Spanish participants. Section 6 presents the results of the replications of the experiment in Portugal and Brazil. Section 7 provides a meta-analysis of the aggregated experimental results obtained in the three experiments. Section 8 discusses the results of the whole set of controlled experiments. Section 9 discusses the threats to validity, and, finally, Section 10 concludes this paper and summarizes directions for further work.

2. Value-based development

This section gives an overview on Value-Based Software Engineering (VBSE) and discusses existing works that contextualize and motivate our present work.

2.1. VBSE definition

The International Organization for Standardization (ISO) defines software engineering as *"the systematic application of scientific and technological knowledge, methods, and experience to the design, implementation, testing, and documentation of software to optimize its production, support, and quality"* [13]. While the ISO definition might satisfy without considering value decisions, it must be extended to consider values effectively. For example, the ISO definition excludes economics, management science, cognitive sciences, and humanities from the body of knowledge required to create successful software systems. In contrast, VBSE considers software development as a purposeful activity carried out by people for people, without ignoring the body of knowledge described above [3]. The ISO definition also delimits the software development by technical activities (e.g., design, implementation, and testing). VBSE considers management-oriented activities that have often been considered slightly [3]. Besides, it covers all practices, activities, and phases involved in software development, addressing a wide diversity of decisions about technical problems, business models, software development processes, software services and products, and related management practices [3]. Also, the ISO definition does not explicitly recognize the ultimate goal of software development: ensuring that software systems continue to meet and adapt to evolving human and organizational needs to create value [1, 3]. According to VBSE, it is not enough for software projects to merely meet unilaterally preset schedule, budget, process, and quality objectives. Instead, it is necessary that the resulting services and products persist in increasing the wealth of the stakeholders and optimizing other relevant value objectives of these projects. Thus, the VBSE definition is *"the explicit concern with value in the application of science and mathematics by which the properties of computer software are made useful to people"* [3].

2.2. Major Elements of VBSE

According to Boehm [3], the major elements of VBSE are: (1) *value-based requirements engineering*, embodying principles and practices to identify stakeholders of systems, elicit their value propositions and reconcile these value propositions into a mutually satisfactory set of objectives for the system; (2) *value-based architecture*, comprising the further adjustment of the system objectives with possible architectural solutions; (3) *value-based design and development*, involving techniques to guarantee that the system's objectives and value considerations are aligned with the business, then inherited by the software design and development practices; (4) *value-based verification and validation*, including both techniques to verify and validate that the software solution satisfies its value objectives; (5) *value-based planning and control*, covering principles and practices to control costs, schedule, and product planning; (6) *value-based risk management*, combining principles and practices to identify, analyze, prioritize and mitigate risk; (7) *value-based quality management*, involving prioritization of desired quality factors concerning according to the stakeholders' value propositions; (8) *value-based people management*, including expectations management, as well as managing the project's accommodation of all stakeholders' value propositions; (9) *value-based software engineering theory*, combining the traditional computer science theories with value-based theories (e.g., utility theory, decision theory, dependency theory, and control theory) to provide processes and framework for guiding VBSE activities.

Our work contributes to the first item of this list, comparing two value-based requirements engineering methods aimed at modeling business values: e3value and DVD. These methods are discussed in Section 3.

3. Selected value-driven modeling methods

This section provides an overview of the two value-driven methods selected: e3value and DVD. Although these methods share the same objective and basic fundamental concepts, their notations, structure, and construction process are significantly different. The section finishes discussing existing studies on comparing value-driven methods.

3.1. The e3value method

The e3value method is, according to Gordijn [14], composed of fifteen concepts and is summarized in [11]. In this work, we use the e3value metamodel introduced in [14] and the concepts introduced in [7]. Figure 1 shows an e3value model with the main concepts, describing an example of a store that sells goods bought from a wholesaler to shoppers.

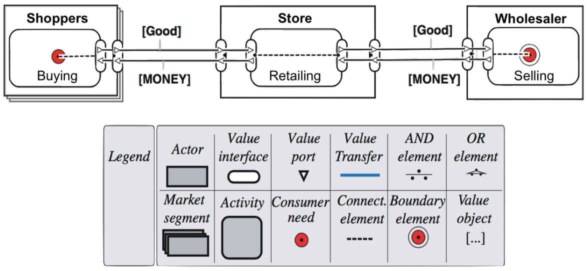

Figure 1: Example of e3value.

The authors define actors, which may be elementary or composite, as environment entities that are economically independent of each other. A composite actor is a group actor with value interfaces of the inner elementary actors. Value interfaces group value ports that provide or request value objects to or from actors or market segments. A market segment is a group of actors that share a set of common properties. A set of value objects defines a value exchange. Value transfers link two value ports. Value transactions are groups of value transfers. For a value exchange to take place, actors, or market segments, must perform a set of operational activities. The collection of these activities is called a value activity. e3value represents value exchange flows (or value stream), inherited from Use Case Maps (UCM) [15], the start (Consumer need) and stop (Boundary element) stimuli, the AND operators, OR operators and connect elements [14]. These concepts are absent in the respective metamodel [14]. A connection element links a start-stop stimulus to a value interface or links value interfaces of the same actor internally. AND and OR operators are used to split or collapse paths of value flows, reusing start and stop stimuli, and partial flows.

3.2. The DVD method

The Dynamic Value Description (DVD) method builds the DVD model to represent value exchanges. The DVD model is, according to the description of the metamodel introduced in [12], composed of seven main concepts [11] and the concepts of the method introduced in [11, 12]. Figure 2-a shows an example of a DVD model and its main concepts for an example similar to that of 1, but here goods can be purchased quickly by the store from a wholesaler and the shopper makes a secure payment.

Similarly to e3value, actors are economically independent environment entities, the difference being that the focus of the business analyst is on defining the *main actor* (central node of the model), and that the focus changes throughout the specification process. Each time

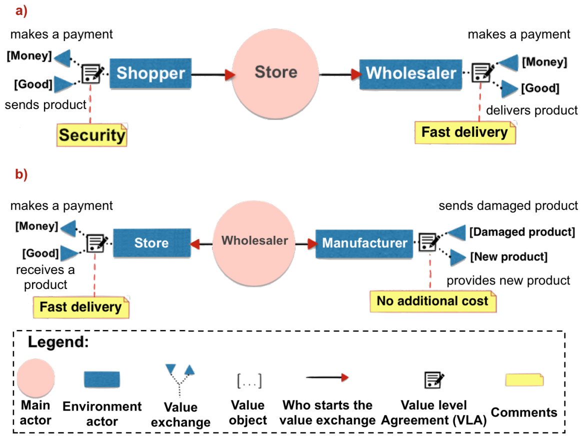

Figure 2: Example of DVD model.

the analyst focuses on one actor (the main actor), identifies new relationships with other *environment actors*, thus producing an inter-organizational network. As the focus changes, new actors and new value exchanges appear, forming new DVD models.

The relationship between the main actor and the environment actor leads to the creation of a *value exchange*. A value exchange shows economic reciprocity through the use of two value ports (arrows connected to value exchange in Figure 2-a), which point to *value objects* (such as money, goods, services). A definition of who starts the value exchanges using a configuration of arrows between the main actor and the environment actors is provided as follows. As the business analyst focuses on one actor, setting the main actor, the supporting tool displays it as the central node of the model, in a dynamic manner. For example, if the business analyst changes the focus of her analysis from the store to the wholesaler, the DVD method sets the wholesaler actor as the main actor (see Figure 2-b) and the business analyst can continue specifying the value exchanges. This second DVD model for wholesaler describes the situation where, if an item of goods is damaged, exchanges can be made directly with the manufacturer at no additional cost. Therefore, all environmental actors can be explored with separate DVD models, and the DVD supporting tool can easily combine all or a subset of the views. Typically, however, we expect that a DVD model is focused on a single main actor, the one representing the business under analysis. Nonetheless, this facility can be used with advantage to model more complex and large organizations. Each value exchange requires a *value level of agreement* between the actors involved, which refers to the minimal business rule or quality of service agreed among them.

3.2.1. Conceptual comparison

Table 1 presents a mapping between e3value and DVD concepts, thus facilitating their comparison.

Table 1: Comparing e3value and DVD concepts.

e3value	DVD
Elementary actor	Main actor or environment actor
Composite actor	Main actor or environment actor
Market segment	Main actor or environment actor
Value interface	Aggregated in value exchange
Value transfer	Aggregated in value exchange
Value port	Value port
Value object	Value object
Value exchange	Value exchange
Value transition	Aggregated in value exchange
Value activity	-
UCM Start stimulus	Who starts
UCM Stop stimulus	Who starts
UCM AND element	Logical operator in value element
UCM OR element	Logical operator in value element
UCM Connect element	-
-	Value level agreement

The e3value model provides more concepts than the DVD model, particularly: value activities, UCM elements, elementary actors, composite actors and market segments. On the other hand, a DVD model aggregates (for simplicity) some of these concepts (see the first three rows in Table 1, for example), and provides the new attribute value level agreement (VLA). A VLA allows a business analyst to specify non-functional requirements or the qualities required for each value exchange. This is important, as the complexity of a software system is determined by its functionality (i.e., what the system does) and global requirements, such as operational costs, performance, reliability, maintainability, portability, robustness [16]. These global requirements, or non-functional requirements (NFR), typically refer to both the operational quality of a system and the constraints imposed on a solution [17]. It is, therefore, possible to define a VLA as an NFR at the business abstraction level.

In summary, while the DVD model is simpler (less concepts), e3value is richer, thus allowing the representation of value streams through the combination of various of its elements. However, despite having fewer concepts, the DVD model represents several of the e3value' concepts but some of these concepts are represented partially or with a different meaning (e.g., UCMs elements). In summary, the DVD model represents the basic concepts found in a value model (e.g., actors, value, and the transfer of values between actors) [18]. The

DVD concepts proved sufficient in the various case studies developed.

3.3. Existing studies comparing value-driven methods

Some comparisons of value modeling methods have been reported, but we are not aware of any existing controlled experiment comparing this kind of methods, with the exception of our previous work [12] that we are extending and complementing in this current paper. The current existing reported studies are rather informal [10, 11, 19, 18, 20].

Kundisch and John [10] compare 12 different business model representations (e.g., BMO, e3value, REA, and others) in relation to their domain of origin (e.g., business, information systems), main concepts, main scope, and design tool. The authors informally compare business representations, not offering an experiment with which to provide empirical evidence regarding which of these methods is the most effective under what circumstances. The work by Gordijn et al. [19] compares BMO and e3value concepts using a framework that maps the similarity of the concepts of the methods. This comparison identifies differences and common characteristics but it is not an experiment to determine which one is better. Andersson et al. [18] compare BMO, REA, and e3value, and identify both a considerable overlap between these methods and their differences. The basic concepts shared among these methods are actors, resources, and the transfer of resources between actors. In another work, Gorgijn et al. [20] compare the iStar and e3value notations to show their complementarity, also aiming to help in the creation, representation, and analysis of e-service business models. Finally, Souza et al. [11] compare the e3value and DVD concepts informally and discuss the design of a planned experiment. Later [12], we show that the DVD model represents most of the business concepts included in the e3value model (DVD additionally includes the notion of value level agreement), and that both value models were built with similar goals in mind.

4. Method for the family of experiments

A family of three experiments was conducted to empirically compare the two value-driven modeling methods selected. The methodology adopted for the experiments is an extension used by Gonzalez-Huerta *et al.* [21] for the five-steps proposed by Ciolkowski *et al.* [22]. The experiments were designed and executed by following the guidelines proposed by Wohlin et al. [23].

Step 1: Experiment preparation. Following the Goal Question Metric (GQM) template [24], the goal of our family of experiments is to **analyze** DVD and e3value models and their modeling processes **for the purpose of** comparing them **with respect to** their actual efficacy (effectiveness and efficiency), perceived efficacy (perceived ease of use and perceived usefulness), and intention to use **in order to** obtain high-quality value models **from the point of view of** both business analysts and software engineers, **in the context of** business modelers (novice software engineers and business analysts).

Step 2: Context definition. The context of the set of experiments is the quality evaluation of two business models carried out by business modelers. The context is defined by (i) the business model to be evaluated, (ii) the value-driven modeling method, and (iii) the selection of participants. Details on the above are provided in Section 5.

Step 3: Experimental tasks. The experimental tasks were structured to allow the comparison of both methods. Depending on the method, each modeling task was composed of the method activities that help to achieve its purpose (e.g., defining a value model). After applying the method, the participants had to fill in a post-experimental questionnaire containing subjective questions regarding their perceptions of the method (see details in Section 5.3).

Step 4: Individual experiments. The family of experiments is summarized in Figure 3. A baseline experiment (UPV) [12] was conducted in Spain. It was first internally replicated in Portugal (UNL) and later externally replicated in Brazil (UFPE), in order to attain more evidence for the results obtained after carrying out the baseline experiment (UPV). The external replications allowed us to increase the external validity.

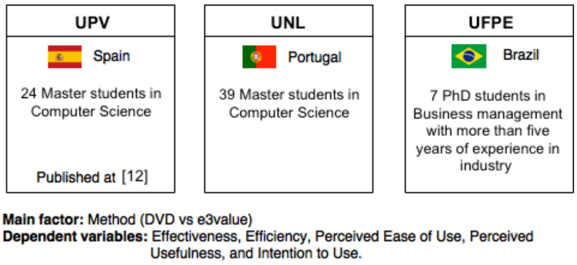
UPV
Spain
24 Master students in Computer Science
Published at [12]
UNL
Portugal
39 Master students in Computer Science
UFPE
Brazil
7 PhD students in Business management with more than five years of experience in industry
Main factor: Method (DVD vs e3value)
Dependent variables: Effectiveness, Efficiency, Perceived Ease of Use, Perceived Usefulness, and Intention to Use.

Figure 3: Overview of our family of experiments.

Step 5: Individual data analysis and meta-analysis. The results of each individual experiment were collected using a spreadsheet and imported to the *SPSS v20* statistical tool [25], after which they were analyzed individually. We then joined the data and imported it to the *R Studio* tool [26] in order to perform the meta-analysis. The analysis procedure is detailed in Section 5.6.

5. Baseline Experiment

An initial design for this experiment was presented and discussed in a workshop held at the ACM/IEEE MODELS conference [11]. The feedback received during the discussion (for example, about the process used to measure the effectiveness of the participants) was taken into account and incorporated into the baseline experiment. The following subsections define the research questions and hypotheses, the sample and participants, the experimental objects and tasks, the metrics and design, and, finally, the analysis procedure of the experiment.

5.1. Research questions and hypotheses

There are two, the research questions addressed by the family of experiments:

RQ1 *Which of the methods has the higher actual efficacy, e3value or DVD?*

RQ2 *Is the perceived efficacy and intention to use of the participants favoring e3value or DVD?*

The independent variable of interest is the use of each value-driven modeling approach with nominal values: DVD and e3value. Two treatments were, therefore, employed in the experiment: the creation of a value model for two software systems using the DVD method and the creation of a value model for the same systems using the e3value method. The experimental data collected made it possible to compare the effects of both treatments. Figure 4 shows the taxonomy of the types of dependent variables used in this experiment.

There are two types of dependent variables in which the treatments are compared: performance-based and perception-based. Performance-based variables assess how well the participants perform the experimental task. They are used to evaluate the actual efficacy of the methods. Perception-based variables assess the participants' perceptions of their performance and their subsequent intention to use the methods DVD or e3value. These variables are used to evaluate the perceived efficacy of the methods, and their likely adoption in practice. There are two performance-based variables:

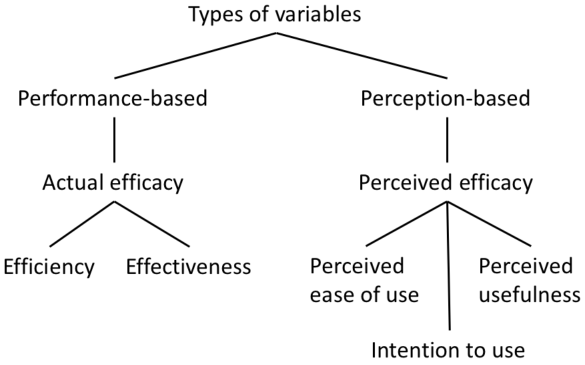

Figure 4: Taxonomy of dependent variables.

- *Efficiency*, measuring the modeling time (i.e., the time required to apply the method).
- *Effectiveness*, measuring the correctness and completeness of the value model created by the participants.

There are also three perception-based variables: *Perceived Ease of Use*, *Perceived Usefulness*, and *Intention to Use*. These variables were identified using the Technology Acceptance Model (TAM) [27], a widely applied theoretical model used to analyze user acceptance and usage behavior as regards emerging information technologies through the use of empirical validations and replications [28]. The perceived efficacy [27] of the method can, therefore, be decomposed into the following subjective dependent variables:

- *Perceived Ease of Use (PEOU)*, indicating the degree to which a person believes that learning and using a particular value-driven method would occur with reduced effort.
- *Perceived Usefulness (PU)*, indicating the degree to which a person believes that using a particular method will increase her/his job performance within an organization.
- *Intention to Use (ITU)*, indicating the extent to which a person intends to use a particular method. It represents a perceptual judgment of the method's efficacy, that is, whether it is cost-effective and is commonly used to predict the likelihood of acceptance of a method in practice.

We formulated several null hypotheses, which were defined in a one-tailed manner since we wished to analyze the effect of the use of value-driven methods on the

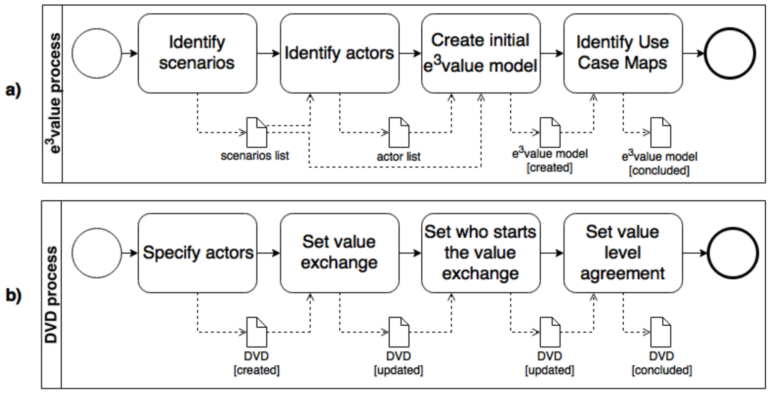

Figure 5: Processes for the creation of (a) an e3value model and (b) a DVD model.

described variables. Each null hypothesis and its alternative are presented as follows:

- H1-0: There is no significant difference between the effectiveness of the DVD and e3value methods / H1-a: The DVD method is significantly more effective than the e3value method.

- H2-0: There is no significant difference between the efficiency of the DVD and e3value methods / H2-a: The DVD method is significantly more efficient than the e3value method.

- H3-0: There is no significant difference between the perceived ease of use of evaluators applying the DVD and e3value methods / H3-a: The DVD method is perceived as easier to use than the e3value method.

- H4-0: There is no significant difference between the perceived usefulness of the DVD and e3value methods / H4-a: The DVD method is perceived as more useful than the e3value method.

- H5-0: There is no significant difference between the intention to use the DVD and e3value methods / H5-a: The DVD method is perceived as more likely to be used than the e3value method.

Although we have no reason to believe that one method is better than the other, the formulation of the hypothesis starts with the DVD method by chance, and we could have chosen the e3value method to start those formulations.

5.2. *Sample and participants*

The sample in the baseline study is a group of 24 MSc students at the UPV, in Spain. The experiment was a class exercise on an Empirical Software Engineering course, which included an introduction to the e3value and DVD methods. The participants had no previous experience of value-driven modeling methods before attending this course. However, they had previous experience in modeling software with UML and had an average of three years experience in software development.

5.3. *Experimental objects and tasks*

Two experimental objects were selected from the following two software requirements systems in literature [29, 30]:

- Wireless access provisioning (Object1): a hotel offers wireless connectivity to businessmen as an additional service.

- Waste management (Object2): waste is traded between an exporter and an importer. The exporter usually pays the importer for the waste handling, but in some cases, the waste can be traded like a regular good, such as recycled waste.

The size of these two experimental objects are comparable. The experimental task was to create a value model following the specific steps of each method (e3value or DVD). Figure 5-a shows the process employed to create an e3value model [14]. Participants had to identify a list of scenarios (or short textual descriptions of the product, service, or experience expected by a customer), after which they had to identify the actors (who offers and who receives the product, service or experience expected) from the list of scenarios. They then had to create the initial e3value model using the products and services mentioned in the list of scenarios and the actors in the list of actors, and add the macro activities in order to operationalize the value exchange. Finally, they had to insert the UCM elements representing the paths of all the scenarios.

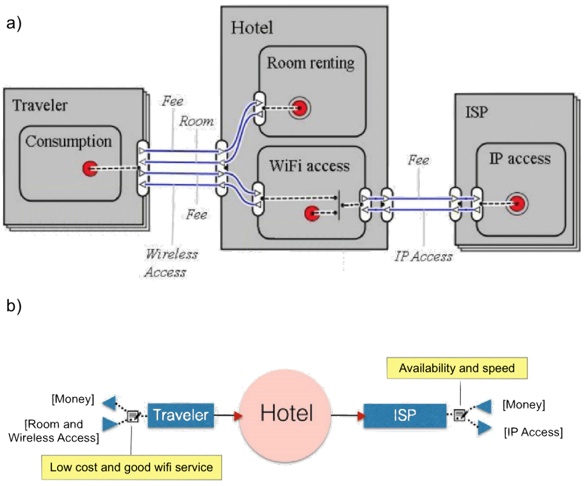

Figure 6: Initial oracles of (a) e3value and (b) DVD for Object1.

Similarly, Figure 5-b shows the process of creating a DVD model. It starts by describing the main actors (the focus of their analysis) and their related environment actors (the model is created like a mind map). This "main actor" focus signifies that the participants have to create as many DVD models as necessary to represent the whole business, after which they must add the value exchanges to the model, define the value elements related to each value port and continue by determining which actor originates the value exchange, checking whether the value elements are specified in the correct value port. The final step is to define the criteria required for value exchanges to be performed and it is crucial to understand the business constraints related to each value exchange.

The correct answers for each of the experimental objects are easily modeled in both DVD and e3value methods. For example, Figure 6 shows the correct answers obtained for the methods e3value (a) and DVD (b) in the case of Object1. These correct answers are used as a baseline (Oracle) to measure the models created by the participants (details in subsection 5.4).

Once the value model had been created, the participants answered the post-experimental questionnaire [31]. This questionnaire, defined as a Google Form, contains a set of closed-questions, allowing the participants to express their opinion on the ease of use, usefulness, and intention to use the method in the future. It also includes three open questions with which to obtain the participants' feedback regarding the changes they would make to improve the method and their reasons for using a given method in the future (if any). The data collected was kept anonymous.

The answers to this questionnaire were the basis used to evaluate the perception-based variables (PEOU, PU, and ITU). The performance-based variables (effectiveness and efficiency) were evaluated by comparing the value model created by the participants with the value model designed by experts and by analyzing the time required to perform each experimental step.

5.4. Metrics

We used an approach based on the information retrieval theory [32] to obtain a quantitative assessment of the Effectiveness of value models modeled with both the e3value and DVD methods. This same approach has been applied in other software engineering experiments [33, 34] to compare models created by participants with an Oracle (the correct model created by an expert) regarding each type of graphic elements through the use of equations (1) and (2) respectively, for *precision* and *recall*, in which the precision$_{\text{element}}$ measures the correctness of a graphical element belonging to a given value model and the recall$_{\text{element}}$ measures the completeness of a value model as regards its graphical element.

$$precision_{element} = \frac{|P_{element} \cap O_{element}|}{|P_{element}|} \tag{1}$$

$$recall_{element} = \frac{|P_{element} \cap O_{element}|}{|O_{element}|} \tag{2}$$

Accordingly, P$_{\text{element}}$ indicates all the particular graphical elements modeled by a participant and O$_{\text{element}}$ represents the known correct set of expected types of graphical elements that can easily be derived using an Oracle.

Precision and recall quantitatively summarize two different concepts. We therefore used their harmonic mean [32] to obtain a balance between the correctness and completeness of each graphical element in a value model (equation 3):

$$F{-}Measure = \frac{2 * precision_{element} * recall_{element}}{precision_{element} + recall_{element}} \tag{3}$$

The F-Measure quantitatively summarizes the accuracy of a value model as regards its graphical elements and is compared with an Oracle.

The effectiveness dependent variable is computed as the arithmetic mean of the entire F-Measure. All the measures above assume values of between 0 and 1. Whatever the measure is, 0 is the worst value and 1 is the best. With regard to effectiveness, values close to 1 signify that the participants defined value models the were very similar to the Oracle. Conversely, values close to 0 indicate that the models were very different from the Oracle. The effectiveness variable has been defined in order to give the same relevance to the correctness and completeness of value models for all the graphical elements of the value model.

The first Oracle was developed by an expert in value modeling before the experiment (one for each experiment object as can be seen in the Figure 6). In the case of e3value, the first Oracle was extracted from literature [29], [30]. As value models could have different levels of granularity, the expert developed new Oracles with different levels of abstraction. For example, in the Oracle represented in Figure 6-a, the participants could create only one activity to represent all hotel services (e.g., "hotel services" activity within the Hotel actor rather than creating the "Room resting" and "WIFI access" activities). At the end, we checked the effectiveness of all models created by the participants against the Oracles, and the higher effectiveness result was selected.

The three subjective variables (e.g., PEOU, PU, and ITU) were measured using a 5-point Likert scale questionnaire with a set of 12 closed-questions: 5 questions for perceived ease of use (PEOU), 5 for perceived usefulness (PU), and 2 for intention to use (ITU) [31]. These were formulated using the opposing statement format, signifying that each question contains two contradictory statements representing the maximum and minimum possible values (5 and 1), where 3 is considered to be a neutral perception. The aggregated value of each variable was calculated as the arithmetical mean of the answers to the questions associated with each perception-based variable. We used Cronbach's alpha test to evaluate the reliability of the survey and of each variable.

5.5. Design and execution

The experiment was planned as a balanced within-participant design with a confounding effect, i.e., the same participants would apply both methods with both experimental objects in a different order. We formed two groups (each of which used one method to one experimental object) to which the participants were randomly assigned. Table 2 summarizes the design of the experiment. The within-participant experimental design is intended to minimize the impact of learning effects on the results since none of the participants repeat any treatment or experimental object during the execution. The comprehension of the software systems requirements may also have affect the application of both methods. We alleviated the influence of this factor by selecting two representative software systems with requirements of a complexity suitable for application in the time slot available for the execution of the experiments (2 sessions of one hour each).

Table 2: Experiment design

Groups	Session 1	Session 2
A	Object1, e3value	Object2, DVD
B	Object1, DVD	Object2, e3value

We conducted a pilot experiment with 2 professors and 1 Computer Science Master's Degree student at the UPV. They played no further part in the controlled experiments. The goals of this pilot experiment were to evaluate all the experimental material, the instructions regarding the experimental procedure and the task completion time. The results indicated that the experiment objects were well suited and that one hour were sufficient to accomplish the task. No software tool was used during the execution of the experiments, to avoid possible usability bias.

A training session explaining the concepts and processes was provided to the participants, who had to create a value model by following the experimental procedure. During the experiment session, the participants were given a pencil, an eraser, sheets of paper and the printed copy of the experimental material slides introducing business modeling and value-driven development, slides describing the value-driven development method and an application example, slides describing the e3value and DVD methods with an example, the specification documents of the software systems to be used in the tasks, and the post-experimental questionnaire. The material was in the participants' native language (e.g., Spanish). No interaction among participants was allowed and no time limit by which the tasks had to be completed was imposed. Moreover, we provided no details on how to deal with the modeling tasks, but any issues concerning the specification documents were clarified. Finally, the participants were asked to register their start and end times for each step performed. The answers to this questionnaire were the basis employed to evaluate the perception-based variables (perceived ease of use, perceived usefulness, and intention to use).

The performance-based variables (effectiveness and efficiency) were evaluated by comparing the value model they created with the value model designed by the expert and by analyzing the time required to perform each experimental step.

5.6. Analysis procedure

We chose to analyze the data collected with statistical tests owing to their robustness and sensitivity and because they have been used in similar experiments ([35], [34]). As is usual, we accepted a probability of 5% of committing a Type-I-Error [23] in all the tests, i.e., rejecting the null hypothesis when it is true. We tested the normality of the data distribution by applying the Shapiro-Wilk test. The results of the normality test allowed us to select the correct significance test with which to examine our hypotheses. When data was assumed to be normally distributed (p-value>0.05), we applied the parametric one-tailed t-test for independent samples [36]. However, when data did not assume the normal distribution (p-value<0.05), we applied the non-parametric Mann–Whitney test [37].

5.7. Summary of Results

The results obtained in the baseline experiment show that the values for all variables are higher for the DVD method (see Table 3). Before applying the analysis procedure (Section 5.6) in order to confirm the results, we used the Cronbach's alpha to examine the reliability of the questionnaire. The test result for Cronbach's alpha for the whole questionnaire was 0.928 and that for each variable was 0.889 (PEOU), 0.802 (PU), and 0.850 (ITU), signifying that the questionnaire is very reliable (i.e., Cronbach's alpha is higher than 0.7 [38]), indicating that the questionnaire is not biased as regards the perceived-based variables.

Figure 7 shows the analysis procedure used to confirm the results. We first applied the Shapiro-Wilk test to verify the normality of the distribution of all variables (effectiveness=0.108, efficiency=0.058, PEOU=0.000, PU=0.465, and ITU=0.005). The results show that effectiveness, efficiency, and PU have a normal distribution (*p-value*>0.05). We therefore applied a *t-test* (parametric test) to verify hypotheses H1-0 (effectiveness), H2-0 (efficiency), and H4-0 (PU) and a Mann-Whitney test (non-parametric test) to check hypotheses H3-0 (PEOU) and H5-0 (ITU).

The *p-value* results obtained from the *t-test* were effectiveness=0.001, efficiency=0.001, and PU=0.121. As the *p-value* for PU is higher than 0.05, we can confirm hypothesis H4-0, meaning there is no significant difference between the methods. Null hypotheses H1-0 and H2-0 must, however, be rejected because the *p-values* for effectiveness and efficiency are lower than 0.05. With regard to PEOU and ITU, the results for the Mann-Whitney test were 0.000 and 0.031, respectively. As both results are lower than 0.05, we cannot confirm hypotheses H3-0 and H5-0, showing that the participants perceived the DVD method to be easier to use than the e3value method (thus confirming H3-a) and their intention to use DVD in the future is higher than that of using e3value (thus confirming H5-a). In summary (see Figure 7), only the result obtained for PU (H4-0) confirms the null hypothesis (artifact in red).

With regard to the RQ1 (Which of the methods has the higher actual efficacy, e3value or DVD?), the data

Table 3: Descriptive statistics for effectiveness, efficiency, PEOU, PU, and ITU per experiment and method.

Experiment	Number of observations	Variable	e3value					DVD				
			Min.	Max.	Med.	Mean	Std. Dev.	Min.	Max.	Med.	Mean	Std. Dev.
UPV	48	Effectiveness	0.26	0.75	0.55	0.56	0.11	0.50	1	0.87	0.83	0.14
		Efficiency	15	56	30.50	33.08	10.85	6	37	16.50	20.04	9.89
		PEOU	1.6	5	3.40	3.41	0.87	1.2	5	4.70	4.25	0.99
		PU	1.8	5	3.40	3.29	0.66	1.6	5	3.80	3.66	0.95
		ITU	1	5	3.25	3.10	1.09	1	5	4.00	3.75	0.96
UNL	78	Effectiveness	0.22	0.82	0.55	0.54	0.14	0.14	1	0.81	0.75	0.22
		Efficiency	7	64	25	29.33	15.89	4	49	20	21.72	13.24
		PEOU	2	4.25	3.25	3.10	0.61	3	5	4.25	4.31	0.61
		PU	2.16	4.33	3.33	3.36	0.58	2.83	4.83	3.66	3.76	0.55
		ITU	1.5	4.5	3	3.19	0.74	2	5	3.5	3.73	0.87
UFPE	14	Effectiveness	0.11	0.65	0.43	0.44	0.18	0.33	1	0.83	0.73	0.28
		Efficiency	24	63	30	38.71	16.55	6	42	23	21.57	12.20
		PEOU	1	4.25	3	2.92	1.36	2.75	5	4.75	4.32	0.82
		PU	1	4.2	3.16	2.85	0.99	3.2	5	3.50	4.04	0.79
		ITU	1	5	3.50	3	1.29	3	5	3.50	3.64	0.97

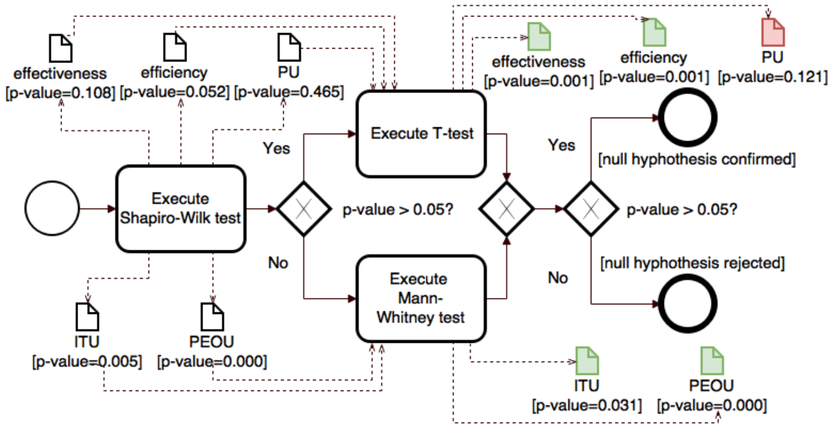

Figure 7: Analysis procedure employed for the experiment in Spain (UPV).

analysis results indicate a significant difference between the methods concerning efficiency (time required to create the model) and effectiveness (correctness and completeness of the model). One plausible justification for this result is that DVD facilitates the representation of the business economic point of view, thanks to its cognitive-based, semi-structured nature. With regard to RQ2 (Is the perceived efficacy and intention to use of the participants favoring e3value or DVD?) the data analysis results show that the perceived efficacy is higher for the DVD method. However, the results show no significant difference between the methods for perceived usefulness (PU). This is not surprising as both methods share the same goal and represent the same central economic concepts. In the case of perceived ease of use (PEOU), the results indicate that the DVD method is significantly easier to use than the e3value method. We also associate this result with the DVD method being being structured as a cognitive mind map.

6. Experimental Replications

This section discusses the experimental replications.

6.1. Motivation

The need to carry out replications (in Portugal and Brazil) of the controlled experiment performed in Spain is justified for two principal reasons. First, the null hypothesis H4-0 (*there is no significant difference between the perceived usefulness of the DVD and e3value methods*) could not be rejected in the (baseline) experiment performed in Spain. This means that the participants perceived both methods to be equally useful for defining value models. As the descriptive statistics analysis shows that the PU result for the DVD method is higher than that of the e3value method, we believed that H4-0 would be rejected if we increased the number of participants. We consequently performed a replication with more participants at the UNL in Portugal. Second, we felt the need to execute an experiment with experienced participants with a business background in order to verify whether the results would hold, thus increasing the validity of the results. This replication was performed at UFPE in Brazil. It is essential to highlight that, with the exception of the experimental material which was translated into the participants' native language (e.g., Portuguese-PT and Portuguese-BR), we did not change any of the experimental conditions of the experiment conducted in Spain. These experiments are, therefore, exact replications of the baseline experiment.

6.2. Sample and participants

The sample in the replication was composed of 46 participants: 39 MSc students in Computer Science in Portugal and 7 Business Management PhD students in Brazil. The 39 MSc students in Computer Science in Portugal were attending the "Software Engineering" and "Requirements Engineering and Software Architecture" courses at Universidade NOVA de Lisboa (UNL). These participants had no previous experience with value-driven modeling methods, but they were experienced in software modeling. In particular, they were familiar with UML and had an average of three years of experience in software development. The experiment took place during April 2017. The 7 Business Management PhD students in Brazil were attending the "Business Process Modeling" course at the Universidade Federal de Pernambuco (UFPE). Before attending this course, these participants had theoretical knowledge of value modeling (e.g., REA [39] and BMO [40]), but no previous experience in the methods used in the experiment. It is worth noting that the Brazilian experiment is important because all the participants are also professionals from industry with more than five years of experience. Despite the small number of participants, the experiment had a balanced within-participant design, what means that the number of observations generated is double the number of participants. The experiment took place during June 2017.

6.3. Results

This section discusses the results from the replications performed in Portugal (UNL) and Brazil (UFPE).

6.3.1. Internal Replication (UNL)

Similarly to the results obtained in the UPV experiment, the descriptive statistics results for all the variables of the UNL experiment also favor the DVD method (see Table 3). Again, Cronbach's alpha was used to examine the reliability of the questionnaire, and the result obtained for the questionnaire was: PEOU questions = 0.803, PU questions = 0.705, ITU questions = 0.732, while that for the whole questionnaire = 0.858. This means that the questionnaire can be considered reliable (Cronbach's alpha is higher than 0.7 [38]).

Figure 8 shows the analysis procedure used to confirm the results of this experiment. After using the process described in the previous section, the Shapiro-Wilk test resulted in the following values for each variable: effectiveness=0.077, efficiency=0.001, PEOU=0.005, PU=0.407, and ITU=0.005). Given that *p-value* > 0.05 for effectiveness and PU, we concluded that the effectiveness and PU data had a normal distribution, so we could apply a parametric statistical test to analyze them. However, it was necessary to apply a non-parametric statistical test to analyze the remaining variables.

We applied a *t-test* to compare the results obtained for effectiveness (0.001) and PU (0.003). These results allowed us to reject hypotheses H1-0 and H4-0 (*p-value*<0.05), meaning that the participants obtained higher quality value models when applying DVD and that they perceived it to be more useful for creating value models than the e3value method.

With regard to the efficiency, PEOU, and ITU variables, the non-parametric test used to compare the results was the Mann-Whitney test. The results were efficiency=0.029, PEOU=0.001, and ITU=0.009. This allowed us to reject hypotheses H2-0, H3-0, and H5-0 because the *p-value*<0.05, signifying that participants created the DVD models significantly faster than the e3value models. The DVD method was also perceived to be considerably easier to use than the e3value method and the participant's intention to use DVD in the future was substantially higher than that of using e3value. Overall, these results confirm that the participants were more efficient and effective when using the DVD method. Unlike the baseline experiment, the results for all the variables in Portugal favored the DVD method, and both research questions (e.g., RQ1 and RQ2) obtained positive responses.

6.3.2. External Replication (UFPE)

The descriptive statistics results obtained for the UFPE experiment show that the DVD method is better ranked in all variables (see Table 3. The results of Cronbach's alpha show that the questionnaire is reliable (all questionnaire=0.952, PEOU questions=0.939,

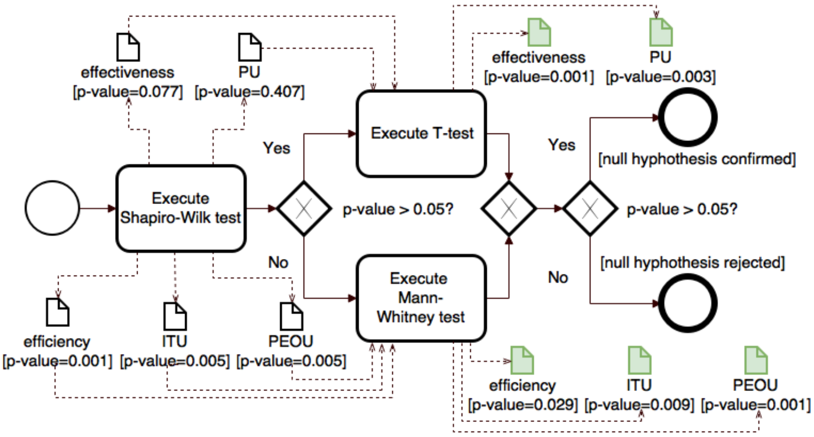

Figure 8: Analysis procedure employed for the experiment in Portugal (UNL).

PU questions=0.920, and ITU questions=0.784), as Cronbach's alpha is higher than 0.7 [38], thus allowing us to apply the analysis procedure. The Shapiro-Wilk test was used to verify the normality of the distribution for each variable. The results (effectiveness=0.618, efficiency=0.263, PEOU=0.048, PU=0.230, and ITU=0.413) show that all the variables have a normal distribution, with the exception of PEOU which has a *p-value*<0.05. The non-parametric Mann-Whitney test was consequently applied in order to verify hypothesis H3-0 (PEOU), while the parametric *t-test* was applied to verify hypotheses H1-0 (effectiveness), H2-0 (efficiency), H4-0 (PU), and H5-0 (ITU). The result of the tests was: effectiveness=0.044, efficiency=0.048, PEOU=0.048, PU=0.030, and ITU=0.316. All the variables, with the exception of ITU, have a *p-value*<0.05, signifying that the null hypotheses can be rejected and confirming the alternative hypotheses H1-a, H2-a, H3-a, and H4-a can be confirmed. In others words, the results show that the participants were more effective and efficient when using the DVD method and they also perceived DVD to be easier to use and more useful than the e3value method. With regard to ITU, as the result obtained from the test was higher than 0.05, we cannot confirm hypothesis H5-a, meaning that there is no significant difference between the participants' intention to use these methods (although the mean value obtained for the DVD method is higher than that obtained for e3value. In summary, the results of this experiment show that DVD was considered better in relation to the variables analyzed, with the exception of ITU.

Figure 9 shows the analysis procedure used to confirm the results of this experiment.

7. Meta-Analysis

Among the existing statistical methods to aggregate results from interrelated experiments [41, 42], meta-analysis allows more general conclusions to be obtained and was, therefore, chosen for this study. Meta-analysis is a set of statistical techniques that can be used to combine and contrast the results (e.g., patterns and sources of disagreement) of multiple scientific studies [43].

Figure 10 shows the forest plot (or blobbogram) provided by the R Studio tool [26] used. The square expresses the magnitude of the effect of the method while the dimensions of the square are proportional to both the weight of the experiment in the meta-analysis and the number of participants. The result for studies with a large sample size is more accurate, meaning that they make a greater contribution to the overall effect [34]. The effect size obtained in our meta-analysis varies between small and medium in all cases. This may indicate that it will be necessary to perform further replications with a larger sample of participants. Despite this, and given that no other similar studies exist in the literature, the present results are still useful and and of interest to the community.

The confidence intervals of each experiment are represented by horizontal lines. We considered a confidence interval of 95 percent for each experiment. When these horizontal lines cross over the central vertical line of the graph, this means that there is no significant difference between the means of the methods (e.g., PU in

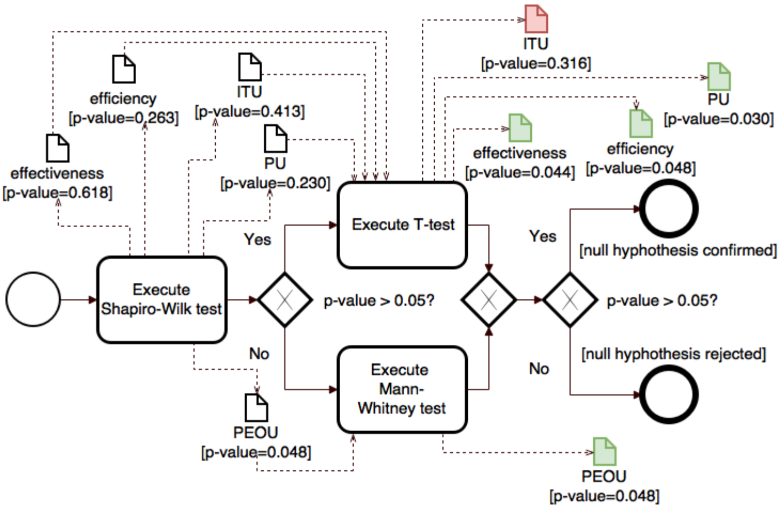

Figure 9: Analysis procedure employed for the experiment in Brazil (UFPE).

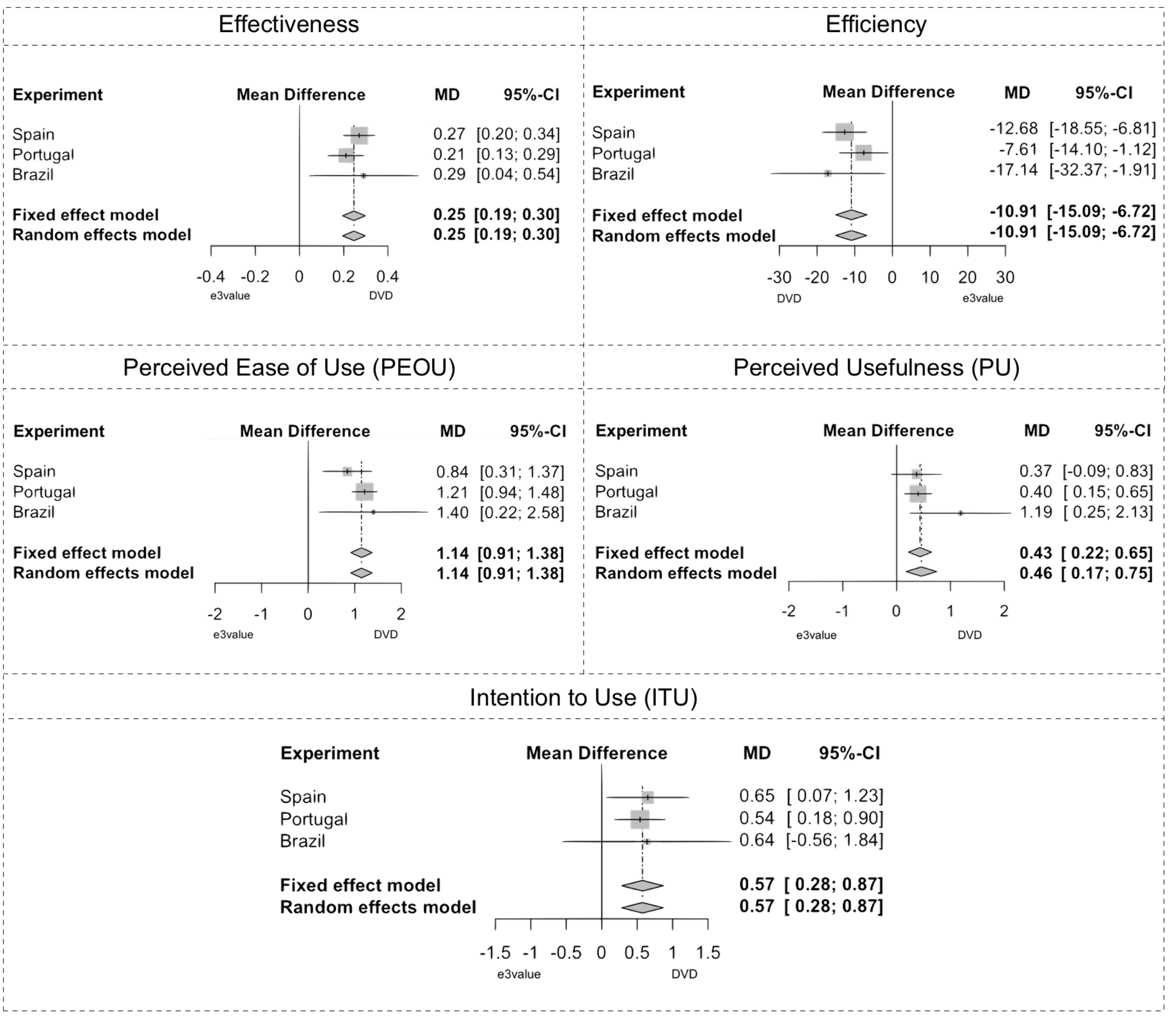

Figure 10: Meta-analysis blobbogram for effectiveness, efficiency, PEOU, PU and ITU.

the experiment conducted in Spain and ITU in the experiment conducted in Brazil). The diamonds represent the overall conclusion. The summary measure is the central line of the diamond, while the associated confidence interval is the lateral tips of the diamond. When the diamond crosses over the central vertical line of the graph, this means that there is no significant difference between the aggregated result. As this did not occur in our meta-analysis, the aggregated result was, therefore, always favorable for one of the methods.

Despite the fact that the null hypotheses H4 (related to PU) and H5 (related to ITU) could not be confirmed in the UPV and UFPE experiments, the overall results of the meta-analysis have a significant positive effect. The diamonds are always positioned on the DVD method side (for example, on the right-hand side of the effectiveness graph) and we can, therefore, reject all null hypotheses. In summary, the meta-analysis strengthens the results obtained in the individual experiments.

8. Discussion of the results

Figure 11 summarizes the descriptive statistic results for the three experiments. The small number of outliers were discarded from the data analysis. These outliers occurred because some participants did not participate in the training session, or arrived late. They just attended the review that was held before each experimental section.

Table 4 summarizes the results for the various hypothesis (where an accepted null hypothesis means no significant difference between e3value and DVD, and

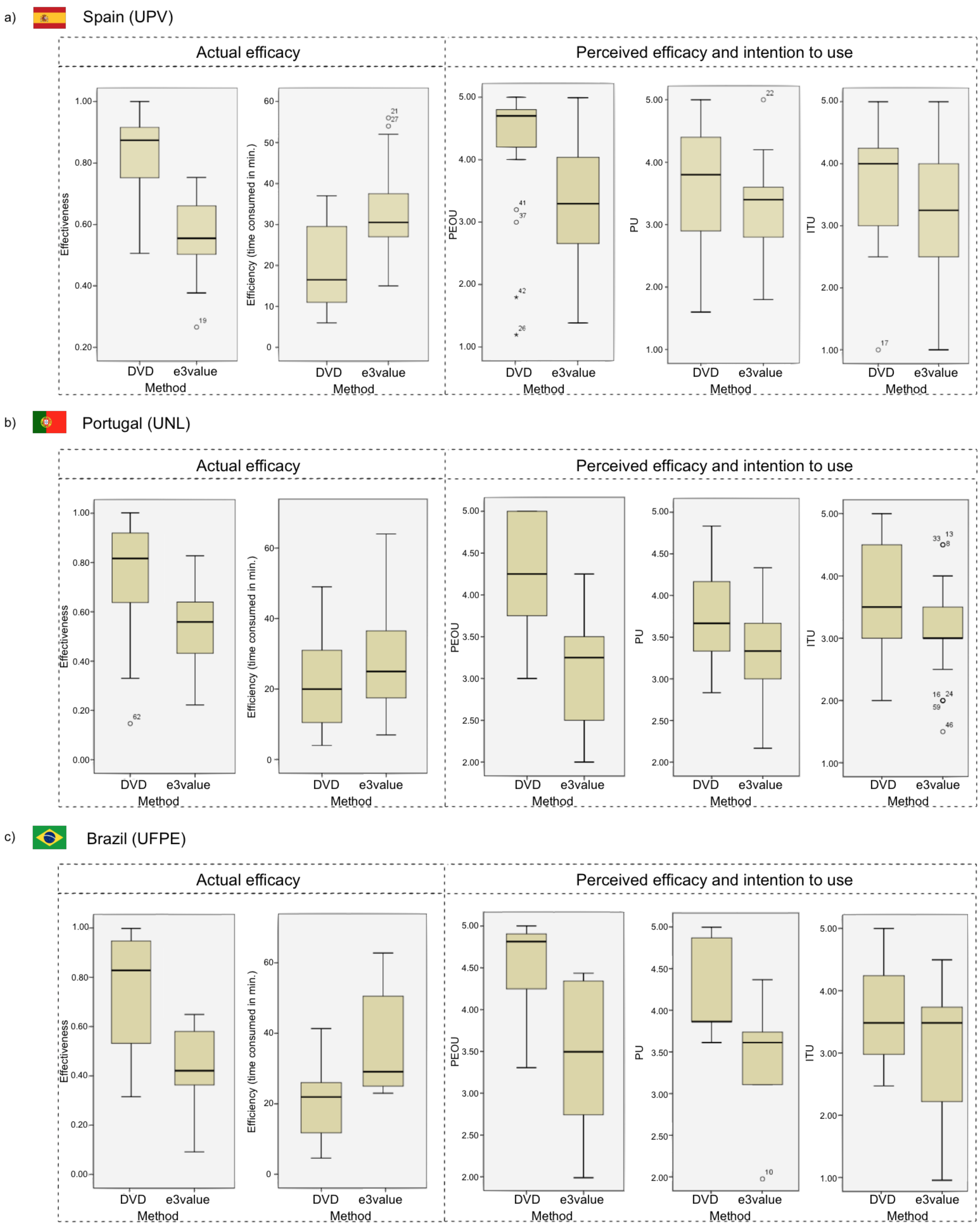

Figure 11: Actual efficacy (effectiveness and efficiency), perceived efficacy (PEOU and PU), and ITU grouped by methods of experiments performed in (a) Spain, (b) Portugal, and (c) Brazil.

an accepted alternative hypothesis means that the result favors DVD). Besides, we calculated the value of *Cohen's d* [44] and the effect-size correlation (*effect size r*) [45] using the means and standard deviations of two groups (treatment and control)[1]. The sign of our *Cohen's d* effect (Cohen's d column in Table 4) indicates the direction of the effect. In the case, the negative sign means that the direction of the effect is in favor of the DVD method. Note that *Cohen's d* result for efficiency is positive, meaning that the participants took longer to model using e3value. In other words, the least efficient method is the one that has the positive result. Regarding the *effect size r*, Cohen provided rules for their interpretation, suggesting that an *effect size r* between |.10| and |.29| represents a "small" effect size, between |.30| and |.49| represents a "moderate" effect size, and larger than |.5| represents a "large" effect size [44]. The last column in Table 4 shows that the effect size of our experiments is, mainly large and moderate. Even though we have some results with small effect size, we believe that the results of our family of experiments are still relevant to the community because there are no other works that empirically compare value-driven development methods (as previously discussed in Section 3.3).

8.1. *Which of the methods has the higher actual efficacy, DVD or e3value? (RQ1)*

The descriptive statistics for effectiveness and efficiency indicate that the DVD method performs better than the e3value method in the experiments performed in Spain, Portugal, and Brazil. The meta-analysis for the aggregated experiments results confirm a significant difference between the methods regarding *efficiency* (time required to create the model) and *effectiveness* (correctness and completeness of the model). One plausible justification for this conclusion is that the DVD method facilitates the representation of the business from an economic point of view, thanks to its cognitive-based, semi-structured nature. Moreover, the DVD method has fewer concepts which might also have a positive effect on the modeling time and the participants' perceived ease of use. The responses for the open questions from the questionnaire indicated that the e3value method has a weak separation of concerns [48]; it represents static (e.g., objects) and dynamic (e.g., scenarios) business concepts in the same model, thus making the value model complex and arduous to build. Furthermore, the participants indicated that *"DVD is very simple, intuitive, and easy to use"*, or *"I would use it [DVD] because it is not difficult to understand and it would be simple to explain to my clients, saving time in modeling this business point of view"*, or *"It [DVD] is not hard to understand and it uses a simple structure"*, or still *"[DVD] makes the business model construction an effective and fast step"*.

In summary, the DVD method appears to represent the essential business value concepts in a structured manner, thus making it a concise technique. The DVD's structure is based on mind map diagrams and inherits the well-known benefits of this structure (e.g., organization, use of keywords, association, grouping ideas, visual memory, and simplicity [49]). The consequence of being concise and having a simple structure seems to help DVD attain more positive results than e3value. (Note that, efficiency in Figure 10 may seem misleading. This is because efficiency is measured in terms of

Table 4: Summary of the results of all experiments, where ✓ means hypothesis accepted and ✗ means hypothesis rejected.

Experiment	Variable	Null hypothesis	Alternative hypothesis	Cohen's d	Effect size r	Effect size interpretation
UPV	Effectiveness	✗	✓	-2.14	-0.73	Large
	Efficiency	✗	✓	1.25	0.53	Large
	PEOU	✗	✓	-0.90	-0.41	Moderate
	PU	✓	✗	-0.45	-0.22	Small
	ITU	✗	✓	-0.63	-0.30	Moderate
UNL	Effectiveness	✗	✓	-1.13	-0.49	Moderate
	Efficiency	✗	✓	0.52	0.25	Small
	PEOU	✗	✓	-1.98	-0.70	Moderate
	PU	✗	✓	-0.70	-0.33	Moderate
	ITU	✗	✓	-0.66	-0.31	Moderate
UFPE	Effectiveness	✗	✓	-1.23	-0.52	Large
	Efficiency	✗	✓	1.17	0.50	Large
	PEOU	✗	✓	-1.24	-0.52	Large
	PU	✗	✓	-1.32	-0.55	Large
	ITU	✓	✗	-0.56	-0.26	Small

[1] Details on how to calculate *Cohen's d* and *effect size r* can be found in [44, 46, 47].

modeling time, meaning that the larger the result, the less efficient the method is, which is why the result for efficiency may appear to be the opposite).

8.2. *Is the perceived efficacy and intention to use of the participants favoring e3value or DVD?*

The result of the descriptive statistics analysis of the replications are in line with the experiment baseline (at UPV). Upon considering the analysis of the hypotheses, the results of the replication contradicts the baseline experiment in relation to PU and ITU. With regard to the PU, we did not confirm a significant difference between DVD and e3value in the UPV experiment (H4-0 was confirmed). However, we believed that H4-0 would be rejected if we increased the number of participants because the analysis of the descriptive statistics in the baseline experiment favors the DVD method. The results of the UNL replication confirmed what we believed, in other words, H4-0 was rejected (the DVD method is perceived as significantly more useful than the e3value method). In the UFPE replication, we changed the participants' background from Computer Science to Business Management. The result for PU in this replication is also favors DVD, and the reason might be that no prior IT knowledge is required to create a DVD model.

The meta-analysis confirmed that, in spite of the result obtained in Spain (UPV), DVD is perceived to be more useful than e3value. One plausible justification for this result is that the DVD method also facilitates the extraction of business knowledge in order to design information systems [50, 51].

One interesting finding that we have identified after carrying out the UFPE replication is that the different backgrounds of the participants (e.g., Computer Science and Business Management) did not significantly alter the results of the experiments. Only the ITU result contradicts those of the other experiments (e.g., UPV and UNL), being significantly higher for for DVD in Spain (UPV) and Portugal (UNL). However, this was not confirmed in Brazil (UFPE), despite the fact that the mean obtained for DVD (3.64) was slightly higher than the one obtained for e3value (3). As the analysis of the descriptive statistics in the UFPE replication shows that the ITU result for DVD is higher than for e3value, we believe that H5-0 (*there is no significant difference between the intention to use the DVD and e3value methods*) would be rejected if we were to increase the number of participants with a business background.

In addition, even when considering the result obtained in Brazil, the aggregated results of the experiments confirm that the participants have the intention to use DVD in the future (when appropriate). Given that the Brazilian participants are practitioners, they suggested that the DVD method needs a supporting tool and integration with business processes (e.g., BPMN [52]) to represent the value stream throughout the business activities. With regard to the integration issue, we would like to emphasize that a DVD model provides a point of view of the business. It needs to be complemented with other models (e.g., process models or goal models) for a more complete representation of the whole business. It is worth highlighting that the DVD method follows a model-driven approach and provides model transformations to the BPMN model [52] (and also to KAOS [53], iStar [54] and SOA services) [50, 51], but this was not part of the experiment. Moreover, even though the e3value method represents value streams (using UCM elements), the result of the descriptive statistics for ITU favored the DVD method.

The questionnaire's open questions also show that the likelihood of intention to use the methods in the future is probably related to the easiness of using the method, as per answers like *"it [DVD] is easier to use and fast to create. Because of this, I would use it in the future"*, or *"I would not use it [the e3value] because it requires a lot of effort to modeling, and provides a complex and confusing diagram. The cost benefit does not pay. I would use it, if it was simpler and more objective"*.

For perceived ease of use (PEOU), results show that DVD is significantly easier than e3value in all the experiments. We also associate this result with the mind map roots of the DVD model. This conclusion is reinforced by the positive answers obtained from the participants' questionnaire, such as, *"(DVD) is easy to understand, simple, and clearly shows those who make the most important exchanges"* and *"I would use this method thanks to its simplicity regarding its use and understanding by non-expert users"*.

Nevertheless, the responses to these open questions also indicated that the participants had some difficulties in understanding the meaning of some modeling elements (e.g., *"[I would like to advise against] using such complicated symbols"*). We plan to perform a new empirical study with the aim of defining a more representative iconography for these methods based on Moody's physics of notation theory [55]. This would be useful as regards improving the visual notation of both methods, thus making them easier to understand and use.

9. Threats to Validity

We must consider certain issues which could threaten the validity of this experiment. With regard to its *inter-*

nal validity, the main threats are: learning effect, fatigue effects, participant experience, information exchange among participants, understandability of the documents, and instrumentation validity. The learning effect was mitigated by ensuring that each group of participants worked with the two methods, on two different experimental objects, using a within-participant experimental design. We mitigated the fatigue effects by executing the experiment in a time slot of 1 hour per session. Regarding the participants' experience, the random heterogeneity of subjects is always present when experimenting with students and we are also conscious that they had no previous knowledge of the value-driven methods being compared. Furthermore, if the knowledge of the students involved in the experiment could be assumed to be comparable to that of junior industry professionals, the working pressure and the overall environment in industry is different. The experiment should be replicated with participants with experience in value-driven modeling. Nevertheless, the experience collected in this first study allows us to refine the material and tasks with the objective of performing a replication in an industrial setting. In order to minimize the information exchange among participants, they were monitored by the experimenters to avoid communication biases while performing the tasks. The understandability of the material was alleviated by performing a pilot study and making it available in three languages (Spanish, Portuguese-PT, and Portuguese-BR). Finally, the selection of different objects in the study may have affected the instrumentation validity and thus biased the results. We mitigated this threat by conducting a pilot experiment to assess both the complexity of the objects and to attempt to identify mistakes in the experimental material.

With regard to *external validity*, the main threats are: representativeness of the results, and the size and complexity of the tasks that might affect the generalization of the results. The representativeness of the results may be affected by the software systems used and the context of the participants selected. We mitigated the selection of software systems by considering a set of artifacts with a similar size and complexity, containing representative artifacts from an existing value-driven development method (i.e., e3value). The size and complexity of the tasks may also affect the external validity. We decided to use relatively small tasks since a controlled experiment requires that participants complete the assigned tasks in a limited amount of time. To confirm or contradict the achieved results, we plan to conduct case studies with larger and more complex tasks.

With regard to *construct validity*, the main threats are: the measures applied in the data analysis and the reliability of the questionnaire. We mitigated this by using measures that are commonly applied in other empirical-based software engineering works (including controlled experiments [23, 34, 56, 57] and meta-analysis [58, 59, 60, 61]). In particular, effectiveness was measured using an information retrieval based approach (see Section 5.1). The subjective variables are based on TAM [27, 28]. The reliability of the questionnaire was tested using the Cronbach test.

With regard to *conclusion validity*, the main threats are: the data collection and the validity of the statistical tests applied. In the case of the data collection, we applied the same data-extraction procedure in each individual experiment and ensured that each dependent variable was calculated with the same formula. With regard to the validity of the statistical tests proposed, we chose those that are most commonly employed in the empirical software engineering field (both for a simple experiment and for meta-analysis) owing to their robustness and sensitivity [62]. Finally, the meta-analysis results may be threatened by the reduced sample size. The effect size for each dataset was found to be small and moderate. To investigate this issue, we plan to conduct further experiment with a large number of participants.

10. Conclusions and future work

This paper reports the results of a family of three controlled experiments carried out to compare two value modeling methods: the widely used e3value method and the Dynamic Value Description (DVD) method. Running a family of experiments rather than an individual experiment provides more evidence about the external validity – including the generalizability – of results [63, 64]. The same hypotheses were tested in three different contexts (Universitat Politècnica de València in Spain, Universidade NOVA de Lisboa in Portugal, and Universidade Federal de Pernambuco in Brazil), employing two different profiles of participants (Master's Degree and PhD students) with two different backgrounds (computer science and business management).

We created sufficient realistic experiment objects for small businesses, and used no support tool to create the corresponding value models, thus avoiding any usability bias. We initially performed the controlled experiment in Spain, and replicated it in Portugal and Brazil with the objective of increasing the evidence and confirming the results obtained with the Spanish participants.

The results show that the efficiency and effectiveness of the DVD method is higher than that of the e3value method as regards representing the business economic point of view in a value model. We also noticed a significant difference among the participants' perceived ease of use, perceived usefulness and intention to use, with the results favoring the DVD method. These results confirm that DVD is a promising method with which to specify business value models.

From a research perspective, the application of the DVD method during the experimental sessions showed us that it could be improved in certain respects (e.g., the elicitation of the VLA and iconography). In addition, owing to its simplicity, we believe that DVD can be used to facilitate the knowledge transfer from the business management area to the information technology area during an information system development.

From a practical perspective, we are aware that this study provides preliminary results on the efficacy of DVD as a business value modeling method. Although the experimental results provided good results, these results need to be interpreted with caution since they are only valid within the context established in this family of experiments. It is now necessary to analyze whether the same results will be obtained with more practitioners and with new experimental objects. Nevertheless, this study has value as a first family of experiments used to evaluate the business value modeling methods with the objective of providing evidence of their efficacy.

Tool support is currently being developed for the DVD method in order to facilitate the value modeling in addition to the generation of the BPMN and goal models using model-driven techniques. In the near future we plan to extend these experiments to explore the how well these methods support knowledge transfer from the business domain to the information systems domain. We additionally plan to carry out a new empirical study to define more a representative iconography in so as improve the effectiveness and ease of use of both methods.

Acknowledgments

We should first like to thank the participants from Spain, Portugal and Brazil who took part in these experiments, for their effort and valuable time. We would also like to thank the Value@Cloud project (MINECO TIN2013-46300-R), the NOVA LINCS Research Laboratory (Ref. UID/CEC/04516/2013), and CAPES – Ciência sem Fronteiras (Ref. 99999.009047/2013-01) for their financial support.

References

[1] S. Biffl, A. Aurum, B. Boehm, H. Erdogmus, P. Grünbacher, Value-based software engineering, Springer Science & Business Media, 2006.

[2] Standish Group, Standish Group Website, http://www.standishgroup.com, 2017.

[3] B. W. Boehm, Value-based software engineering: Overview and agenda, in: Value-based software engineering, Springer, 2006, pp. 3–14.

[4] D. J. Reifer, Making the software business case: Improvement by the numbers, Pearson Education, 2001.

[5] M. Denne, J. Cleland-Huang, Software by numbers: Low-risk, high-return development, Prentice Hall Professional, 2003.

[6] J. Gordijn, H. Akkermans, Designing and evaluating e-business models, IEEE Intelligent Systems 16 (2001) 11–17.

[7] J. Gordijn, E^3*-value* in a Nutshell, International Workshop on e-Business Modeling, HEC Business School (2002).

[8] A. Rasiwasia, Meta Model for Business Model Design: Designing a Meta model for E3 value model based on MOF, Master's thesis, Department of Computer and Systems Sciences, Royal Institute of Technolog. Stockholm, Sweden, 2013.

[9] V. Kartseva, J. Gordijn, Y.-H. Tan, Inter-organisational Controls as Value Objects in Network Organisations, in: Dubois E., Pohl K. (eds) Advanced Information Systems Engineering. CAiSE 2006. Lecture Notes in Computer Science, volume 4001, Springer, Berlin, Heidelberg, 2006.

[10] D. Kundisch, T. John, Business Model Representation Incorporating Real Options: An Extension of e3-Value, in: 45th Hawaii International Conference on System Sciences (HICSS), IEEE, Hawaii, USA, 2012.

[11] E. Souza, S. Abrahao, A. Moreira, J. Araújo, E. Insfran, Comparing Value-Driven Methods: an experiment design, in: HuFaMo'16 held in MODELS, Saint Malo, France, 2016.

[12] E. Souza, S. Abrahao, A. Moreira, E. Insfran, J. Araújo, Evaluating the efficacy of value-driven methods: a controlled experiment, in: 26th International Conference on Information Systems Development (ISD2017 Cyprus), Larnaca, Cyprus, 2017.

[13] B. ISO, Iec 2382-1 1993, Information technology. Vocabulary. Fundamental terms. British Standards Institution (1994).

[14] J. Gordijn, Value-based Requirements Engineering:exploring innovative e-commerce ideas. Vrije Universiteit Amsterdam. Amsterdam, The Netherlands, Ph.D. thesis, 2002.

[15] R. J. Buhr, R. S. Casselman, Use case maps for object-oriented systems, Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1995.

[16] L. Chung, J. C. S. do Prado Leite, On non-functional requirements in software engineering, in: Conceptual modeling: Foundations and applications, Springer, 2009, pp. 363–379.

[17] J. Rumbaugh, I. Jacobson, G. Booch, Unified modeling language reference manual, the, Pearson Higher Education, 2004.

[18] B. Andersson, M. Bergholtz, A. Edirisuriya, T. Ilayperuma, P. Johannesson, J. Gordijn, B. Grégoire, M. Schmitt, E. Dubois, S. Abels, et al., Towards a reference ontology for business models, Conceptual Modeling (ER 2006) (2006) 482–496.

[19] J. Gordijn, A. Osterwalder, Y. Pigneur, Comparing two business model ontologies for designing e-business models and value constellations, BLED (2005).

[20] J. Gordijn, E. Yu, B. van der Raadt, E-service design using i* and e3value modeling, IEEE software, volume 23, issue 3 (2006) 26–33.

[21] J. Gonzalez-Huerta, E. Insfran, S. Abrahão, G. Scanniello, Validating a model-driven software architecture evaluation and improvement method: A family of experiments, Information and Software Technology 57 (2015) 405–429.

[22] S. Biffl, M. Ciolkowski, F. Shull, A family of experiments to investigate the influence of context on the effect of inspection techniques, Proceedings of the Empirical Assessment in Software Engineering, IEE (2002).

[23] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Reg-

nell, A. Wesslén, Experimentation in Software Engineering, Springer-Verlag Berlin Heidelberg, 2012.

[24] V. R. Basili, H. D. Rombach, The Tame Project - Towards Improvement-Oriented Software Environments, IEEE Transactions on Software Engineering, volume 14, issue 6 (1988) 758–773.

[25] IBM Analytics, IBM SPSS Software: Deliver greater business results with Predictive Intelligence, Available on https://goo.gl/kfth4N, 2018.

[26] R. Studio, Rstudio: integrated development environment for r, RStudio Inc, Boston, Massachusetts (2012).

[27] F. D. Davis, Perceived usefulness, perceived ease of use, and user acceptance of information technology, MIS quarterly, Volume 13, Issue 3 (1989) 319–340.

[28] W. R. King, J. He, A meta-analysis of the technology acceptance model, Information & Management, volume 46, issue 6 (2006) 740–755.

[29] C. Huemer, A. Schmidt, H. Werthner, A UML profile for the e3-value e-business model ontology, Proceedings of the Third International Workshop on Business/IT Alignment and Interoperability (BUSITAL'08) held in conjunction with CAiSE'08 Conference, CEUR-WS, Volume 336. Montpellier, France (2008).

[30] Z. Derzsi, J. Gordijn, A Framework for Business/IT Alignment in Networked Value Constellations., T. Latour, & M. Petit (Eds.), 18th International Conference on Advanced Information Systems Engineering (pp. 219-226). Namur, Belgium: Namur University Press. (2006).

[31] E. Souza, S. Abrahao, A. Moreira, J. Araújo, E. Insfran, Value-Driven Development Method Survey Instrument, Available on https://goo.gl/forms/6KJA3zXyUFx3iHa62, 2017.

[32] W. B. Frakes, R. Baeza-Yates, Information Retrieval: Data Structures & Algorithms, Prentice-Hall, Inc., Upper Saddle River, New Jersey, USA, 1992.

[33] G. Scanniello, U. Erra, Distributed modeling of use case diagrams with a method based on think-pair-square: Results from two controlled experiments, Journal of Visual Languages & Computing, volume 25, issue 4 (2014) 494–517.

[34] S. Abrahao, C. Gravino, E. Insfran, G. Scanniello, G. Tortora, Assessing the Effectiveness of Sequence Diagrams in the Comprehension of Functional Requirements: Results from a Family of Five Experiments, IEEE Transactions on Software Engineering, volume 39, issue 3 (2013) 327–342.

[35] L. C. Briand, Y. Labiche, M. Di Penta, H. Yan-Bondoc, An experimental investigation of formality in UML-based development, IEEE Transactions on Software Engineering, volume 31, issue 10 (2005) 833–849.

[36] N. Juristo, A. M. Moreno, Basics of Software Engineering Experimentation, Springer US, 1st edition, 2010.

[37] W. J. Conover, Practical Nonparametric Statistics, Wiley, 3rd edition, 2006.

[38] N. J, Psychometric theory, McGraw-Hill, New York, NY, second ed. edition, 1978.

[39] W. E. McCarthy, The REA accounting model: A generalized framework for accounting systems in a shared data environment, Accounting Review, volume 57, issue 3 (1982) 554–578.

[40] A. Osterwalder, The Business Model Ontology-a proposition in a design science approach (2004), Ph.D. thesis, Universite de Lausanne, 2004.

[41] L. V. Hedges, I. Olkin, Statistical Methods for Meta-Analysis, Academia Press, 1st edition, 1985.

[42] R. Rosenthal, Meta-Analytic Procedures for Social Research, Sage Publications, volume 6, Applied Social Research Methods Series, 1991.

[43] K. J. Rothman, S. Greenland, T. L. Lash, Meta-Analysis. Page 652 in Modern epidemiology, Lippincott Williams & Wilkins, 2008.

[44] J. Cohen, Statistical power analysis for the behavioral sciences 2nd edn, 1988.

[45] J. C. Nunnally, I. H. Bernstein, Psychometric theory (1978).

[46] L. A. Becker, Effect Size Calculators, University of Colorado Colorado Springs. Available on https://www.uccs.edu/ lbecker/, 1999.

[47] L. A. Becker, Effect Size (ES), University of Colorado Colorado Springs. Available on https://www.uccs.edu/lbecker/effect-size.html, 2000.

[48] E. W. Dijkstra, On the role of scientific thought, in: Selected writings on computing: a personal perspective, Springer, 1982, pp. 60–66.

[49] T. Buzan, B. Buzan, The Mind Map Book: How to Use Radiant Thinking to Maximize Your Brain's Untapped Potential, Plume; Reprint edition, 1996.

[50] E. Souza, A. Moreira, Aligning business models with requirements models, in: Themistocleous M., Morabito V. (eds) Information Systems. EMCIS 2017. Lecture Notes in Business Information Processing, volume 299, Springer, Cham, 2017.

[51] E. Souza, A. Moreira, C. De Faveri, An approach to align business and it perspectives during the soa services identification, 17th International Conference on Computational Science and Its Applications, Trieste, Italy (2017).

[52] S. A. White, Introduction to BPMN, IBM Cooperation. Available on https://goo.gl/YcNCKN, 2004.

[53] A. Dardenne, A. van Lamsweerde, S. Fickas, Goal-directed requirements acquisition, Science of Computer Programming, volume 20, issue 1-2 (1993) 3–50.

[54] E. Yu, P. Giorgini, N. Maiden, J. Mylopoulos, Social modeling for requirements engineering, Mit Press, 2011.

[55] D. Moody, The Physics of Notations: Toward a Scientific Basis for Constructing Visual Notations in Software Engineering, IEEE Transactions on Software Engineering, volume 35, issue 6 (2009) 756–779.

[56] V. R. Basili, R. W. Selby, D. H. Hutchens, Experimentation in software engineering, IEEE Transactions on software engineering (1986) 733–743.

[57] S. Abrahao, E. Insfran, J. A. Carsí, M. Genero, Evaluating requirements modeling methods based on user perceptions: A family of experiments, Information Sciences 181 (2011) 3356–3378.

[58] W. Hayes, Research synthesis in software engineering: a case for meta-analysis, in: Software Metrics Symposium, 1999. Proceedings. Sixth International, IEEE, pp. 143–151.

[59] A. J. Sutton, K. R. Abrams, D. R. Jones, An illustrated guide to the methods of meta-analysis, Journal of evaluation in clinical practice 7 (2001) 135–148.

[60] M. Pai, M. McCulloch, J. D. Gorman, N. Pai, W. Enanoria, G. Kennedy, P. Tharyan, J. J. Colford, Systematic reviews and meta-analyses: an illustrated, step-by-step guide., The National medical journal of India 17 (2004) 86–95.

[61] M. W. Lipsey, D. B. Wilson, Practical meta-analysis., Sage Publications, Inc, 2001.

[62] K. Maxwell, Applied statistics for software managers, Prentice Hall, 2002.

[63] V. R. Basili, F. Shull, F. Lanubile, Building knowledge through families of experiments, IEEE Transactions on Software Engineering, volume 25, issue 4 (1999) 456–473.

[64] J. C. Carver, N. Juristo, M. T. Baldassarre, S. Vegas, Replications of software engineering experiments, Empirical Software Engineering 19 (2014) 267–276.