Evaluating Requirements Methods based on User Perceptions: A Family of Experiments

Silvia Abrahão[1], Emilio Insfran[1], José Angel Carsí[1], Marcela Genero[2]

[1]ISSI Research Group, Department of Information Systems and Computation, Universidad Politécnica de Valencia, Camino de Vera, s/n, 46022, Valencia, Spain, Phone: +34 96 387 7000 – ext. 83510, Fax: +34 96 387 7359, E-mail: {sabrahao, einsfran, pcarsi}@dsic.upv.es

[2]ALARCOS Research Group, Department of Technologies and Information Systems, University of Castilla-La-Mancha, Paseo de la Universidad Nº 4, 13071, Ciudad Real, Spain Phone: 926 295300 – ext.3747, Fax: +34 926 295354, E-mail: Marcela.Genero@uclm.es

ABSTRACT

Numerous methods and techniques have been proposed for requirements modeling, although very few have had widespread use in practice. One drawback of requirements methods is that they lack proper empirical evaluations. This means that there is a need for evaluation methods that consider both the theoretical and practical aspects of this type of methods and techniques. In this paper, we present a method for evaluating the quality of requirements modeling methods based on user perceptions. The method consists of a theoretical model that explains the relevant dimensions of quality for requirements methods, along with a practical instrument with which to measure these quality dimensions Basically, it allows us to predict the acceptance of a particular requirements method in practice, based on the effort of applying the method, the quality of the requirements artifacts produced, and the user perceptions with regard to the quality of the method. The paper also presents an empirical test of the method that has been proposed for evaluating a Rational Unified Process (RUP) extension for requirements modelling. That test was carried out through a family of experiments conducted with students and practitioners and provides evidence of the usefulness of the evaluation method proposed.

Keywords: Requirements Modeling, Method Evaluation, RUP, Theoretical Model, Controlled Experiment.

1. INTRODUCTION

The Requirements Engineering (RE) process is recognized as being the most critical process in software development. Errors made during this process may have negative effects on subsequent development steps, as well as on the quality of the resulting software. Research into the requirements engineering process has produced an extensive body of knowledge, along with different types of methods, notations, and automated tools, in most cases.

Despite the existence of numerous methods for requirements modeling, very few of these have had widespread use in practice. One drawback of requirements modeling methods is that they lack proper empirical evaluations. As pointed out by Cheng and Atlee [9], the ultimate impact of requirements engineering research depends on how relevant the results are to industry. However, if practitioners are to consider adopting a given requirements modeling method, they must know how effective it is, as well as how it compares with other similar methods.

Despite the efforts already made to define approaches for evaluating requirements modeling methods, we believe that an evaluation method that considers both theoretical and practical aspects of this type of methods has yet to be developed. We attempt to address this issue by proposing a method with which to evaluate the quality of requirements modeling methods based on user perceptions. This method consists of a theoretical model which explains the relevant dimensions of quality for requirements modeling methods, together with a practical instrument to measure these quality dimensions.

We use existing theories and models to determine the appropriate dimensions (and their relationships) for evaluating requirements modeling methods from the user point of view. To be more specific, we adapt the Method Evaluation Model (MEM) [28], which is a theoretical model for evaluating Information Systems (IS) design methods. The MEM incorporates both aspects of method “success”: actual performance and likelihood of acceptance in practice. It combines Rescher’s Theory of Pragmatic Justification [32], a theory for validating methodological knowledge, with Davis’s Technology Acceptance Model (TAM) [13]. Our evaluation model allows us to predict the likelihood of a particular method being accepted in practice, based on the effort of applying the method, the quality of the produced requirements, and the user perceptions of the quality of the method.

To test the usefulness of the proposed evaluation method (i.e., the adapted MEM) empirically, we performed a family of four controlled experiments with students and practitioners, in order to evaluate a Rational Unified Process (RUP) extension for requirements modeling [19] [20]. A family of experiments contains multiple similar empirical studies that pursue the same goal. As Basili *et al.* [5] observe, and as we have corroborated in previous research [11] [33], a family of experiments builds the knowledge that is needed to extract significant conclusions that can be applied in practice. That being so, the primary goal of the family was to test the usefulness of MEM in the evaluation of requirements modeling methods. To attain this goal, we pursued another objective, which was to evaluate the likelihood of acceptance in practice of a specific requirements modeling method, i.e., the Rational Unified Process (RUP) extension.

This paper is organized as follows. Section 2 discusses existing methods and models for evaluating requirements modeling methods. In Section 3 the MEM is described, along with its adaptation for use with requirements modeling methods. Section 4 introduces the RUP extension for modeling requirements which is used to test our proposed evaluation method. An overview of the family of experiments conducted to empirically test the proposed evaluation method is set out in Section 5, and this is followed by a description of the design of the individual experiments in Section 6. Section 7 presents the individual data analysis and Section 8 gives a summary of the results of the experiments. It also discusses the limitations of the evaluation method proposed, along with the limitations of the empirical tests conducted to validate the proposed evaluation method. Finally, Section 9 summarizes the conclusions, as well as the lessons learned for future work.

2. RELATED WORK

As a young, multi-disciplinary field, Requirements Engineering (RE) still lacks a broad consensus on appropriate research methodologies and evaluation criteria. One of the first attempts to identify and measure the quality of software requirements was presented by Davis *et al.* [12]. Most of the evaluation criteria, however, were presented in the form of unsystematic lists of the properties that were desired. As far as the quality evaluation of the requirements modeling methods is concerned, several approaches have been proposed here. For example, Al-Subaie and Maibaum [3] evaluated a goal-oriented requirements engineering method (KAOS) and its support tool (Objectiver) with respect to a number of general RE issues such as traceability, validity, and completeness. The purpose of this study was to investigate the range of success of the KAOS method in dealing with these RE difficulties when it was applied to solve a target problem. This experience was used to evaluate the strengths and weakness of KAOS, as well as to explain why the KAOS method and the Objectiver tool are/are not appropriate for helping to solve the difficulties of RE.

In a similar study, Geisser *et al.* [15] presented an evaluation method for requirements engineering approaches in distributed software development projects. They developed an adaptable and cost-effective evaluation method for distributed RE methods, in addition to the corresponding tool for conducting evaluation projects, both in terms of RE process efficiency and effectiveness (i.e., quality of specification).

Nikula and Sajanie [29] put forward an evaluation framework that is applied to evaluate a new domain-specific method which has been designed to ease the adoption of basic requirements engineering practices in small organizations. This framework, which is based on discovering studies dealing with the adoption of technological innovations, uses contextual factors and method-specific factors to guide the evaluation efforts made to detect causalities.

Although several frameworks for evaluating requirements modeling methods have been proposed, very few have had widespread use in practice. One drawback of these approaches is that they lack theoretical foundations [7]. This deficiency has in turn brought about the appearance of several theories and models whose aim is to evaluate the quality of Information Systems design methods and conceptual models. Most of these theories have their roots in theoretical models defined in the field of Social Sciences.

Such theoretical models incorporate constructs with which to measure the user's psychological reactions and organizational factors systematically. Of all the models that have been proposed for user technology acceptance, the TAM proposed by Davis [13] has been one of the most influential. This model allows us to predict the likelihood of a new technology being accepted and/or adopted within a group or an organization.

In the field of Software Engineering, some theoretical models have been used to explain the acceptance of methodologies and tools on the part of the software developer. For instance, Riemenschneider *et al.* [34] examined five theoretical models of individual intentions to accept tools in the context of software methodologies. In another study, Ali Babar *et al.* [4] extended the TAM model, in order to evaluate user acceptance of a groupware tool for the process of evaluation of software architecture. Despite its wide use, some problems with the TAM model have recently been reported [36], e.g., the *actual* use of the technology, as opposed to the behavioural *intention* to use it, is rarely monitored. That said, this model has been successfully adapted, by means of the MEM, to predict the likelihood of acceptance in practice of Information Systems design methods [28] and of functional size measurement methods [1]. In this paper we therefore propose that the MEM could also be adapted to evaluate the likelihood of requirements modeling methods being accepted in practice.

3. EVALUATION METHOD

The evaluation method proposed in this paper is based on the adaptation of the MEM [28] for its use with requirements modeling methods. The MEM provides mechanisms for evaluating both the *likelihood of acceptance* and the *actual impact* of a method in practice. The likelihood of acceptance is indicated for recently-proposed methods, while the actual impact can only be measured for those which are already established. We believe that the MEM provides a suitable basis for predicting the likelihood of acceptance, as well as for giving an idea of what the actual impact of requirements modeling methods in practice will be. An overview of the MEM is presented below, followed by an explanation of the adaptation we performed.

3.1 The Method Evaluation Model (MEM)

The main contribution of the MEM, in comparison to alternative models, is that it incorporates two different aspects of method success: actual efficacy and actual usage (see Figure 1). This means that the adoption of a method in practice depends not only on whether it is actually effective (pragmatic success), but also on whether the users of the method perceive it to be effective (perceived success). Both aspects must be considered when evaluating requirements modeling methods. Figure 1 shows the constructs of the model, along with the hypothesized causal relationships among the model's constructs.

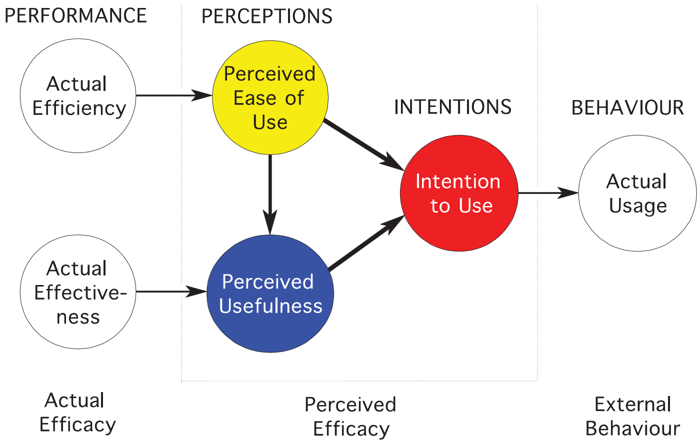


Figure 1. The Method Evaluation Model

In the MEM, efficacy is defined as a separate construct, which is different from efficiency and effectiveness. The *efficacy* construct is derived from Rescher's notion of pragmatic success [32], which is defined as the *efficiency* and *effectiveness* with which a method achieves its objectives. The evaluation of a method's efficacy thus requires the measurement of both the effort required (efficiency) and the quality of the results (effectiveness).

The constructs of the MEM are based on the Technology Acceptance Model (TAM) [13], a well-known and thoroughly validated model for evaluating information technologies. The constructs of the MEM are:

- *Actual Efficacy*, which consists of two performance-based variables:
  - *Actual Efficiency*: the effort required to apply a method.

    - *Actual Effectiveness*: the degree to which a method achieves its objectives. This construct is related to the quality of the artifact(s) obtained by applying the method. According to Rescher, all methods are intended to achieve certain objectives. Rescher defines a method as “a collection of rules and procedures designed to assist people in performing a particular task” [32]. Different types of methods are defined by different objectives. This means that specific dependent variables will need to be defined for each class of methods, in order to measure performance with regard to its specific objectives.

- *Perceived Efficacy*, which consists of two perception-based variables:
    - *Perceived Ease of Use (PEOU)*: the degree to which a person believes that using a particular method would be effort-free. This variable represents a perceptual judgment of the effort required to learn and use a method.
    - *Perceived Usefulness (PU)*: the degree to which a person believes that a particular method will achieve its intended objectives. This variable represents a perceptual judgment of the method’s effectiveness. There is a causal relationship in the model which indicates that perceived usefulness can be determined by perceived ease of use.
- *Intention to Use (ITU)*: the extent to which a person intends to use a particular method. This variable represents a perceptual judgment of the method’s efficacy – that is, whether it is cost-effective. This variable is used to predict the *likelihood of acceptance* of a method in practice. The hypothesized causal relationships suggest that perceived ease of use and perceived usefulness directly affect intentions to use a method.
- *Actual Usage*: a behavior-based variable, defined as the extent to which a method is used in practice (as opposed to potential impact as defined by Intention to Use). This variable is used to measure the *actual impact* of a method in practice. According to the hypothesized causal relationship, actual usage will be determined by intention to use. This relationship was defined in accordance with the empirical studies of the

Technology Acceptance Model, which have reported a highly significant causal link between behavioral intention to use and actual behavior [25].

In the next section we shall present how the constructs of the MEM have been adapted for the evaluation of requirements modeling methods.

3.2 Adapting MEM for its use with requirements modeling methods

The first step involved in adapting the MEM is to define the specific objectives of requirements modeling methods. The general constructs of the MEM can then be instantiated into concrete dependent variables based on these objectives.

We consider that requirements modeling methods have two primary objectives: (1) to capture user and software requirements, in order to facilitate communication among stakeholders and developers, and (2) to define precise specifications of what the software system is intended to do. We distinguish between two types of dependent variables: *performance-based variables*, which measure how well subjects are able to understand or use a requirements modeling method, and *perception-based variables*, which measure how effective subjects believe a requirements modeling method is.

Evaluating the performance of a requirements modeling method involves measuring the effort required (input) and the quality of the requirements modeling artifacts (outputs). The effort required to understand and/or apply the method (i.e., actual efficiency) can be measured by using several measures such as time, cost, and cognitive effort.

The quality of the method’s result (actual effectiveness) can be measured by evaluating the artifacts that are produced by using the requirements modeling method. Actual effectiveness can be measured by using specific quality characteristics and attributes for products, such as those provided by the ISO/IEC 9126 [21] (e.g., reusability, usability, understandability).

In this work, we focus on the understandability of requirements models obtained using a particular requirements modeling method, since this is essential for the validation of requirements between stakeholders and developers. A model must first be comprehended before any desired changes to it can be identified, designed, or implemented. In addition, Gemino and

Wand [16] suggest that the usefulness of any modeling method should be evaluated on the basis of its ability to represent, communicate, and understand the domain.

Some empirical studies suggest that problem-solving tasks can be used as an instrument for measuring understandability (e.g., [16][7]). The answer to a problem-solving question requires the problem solver to reason about the domain that is represented in the model. The following performance-based variables are therefore used to measure subjects' actual efficiency and actual effectiveness in understanding the semantics of requirements models:

- Understandability Effectiveness: this is the average of correct answers divided by the total number of questions for each of the requirements models used. This variable is calculated as follows:

$$UnderstandabilityEffectiveness = \frac{\sum_{i=1}^{n} \left(CorrectAnswers_i \,/\, NumerOfQuestions_i \right)}{n} \quad (1)$$

- Understandability Efficiency: this is the average of correct answers divided by the time spent on providing the answers for each of the requirements models used. This variable is calculated as follows:

$$UnderstandabilityEfficiency = \frac{\sum_{i=1}^{n} \left(UnderstandabilityEffectiveness_i / Time_i \right)}{n} \quad (2)$$

To measure the perception-based variables, we relied on an existing measurement instrument for the MEM, although we adapted this instrument for use with requirements modeling methods (basically by rewording statements).

Figure 2 (a) shows how the MEM was operationalized to evaluate requirements modeling methods. In order to measure each of the three main constructs, Perceived Ease of Use (PEOUi), Perceived Usefulness (PUi), and Intention to Use (ITUi), we defined sets of questions based on the items shown in Table 1. Figure 2 (b) shows the theoretical model proposed to evaluate the quality of requirements modeling methods. Basically, we use the performance-based measures as influencing factors for the perceived-based variables. According to our MEM adaptation, the likelihood of a requirements modeling method being accepted in practice can be predicted by testing the following hypotheses (see Figure 2 (b)):

- $H1_0$: The requirements modeling method is perceived as difficult to use. $H1_1=\neg H1_0$

- $H2_0$: The requirements modeling method is perceived as not useful. $H2_1 = \neg H2_0$
- $H3_0$: There is no intention to use the requirements modeling method in the future. $H3_1 = \neg H3_0$

These hypotheses relate to a direct relationship between the use of a particular requirements modeling method and the users' performance, perceptions, and intentions. The evaluation model also proposes a number of hypotheses that indicate causal links between dependent variables (such as performance having an effect on perceptions or perceptions influencing intentions):

- $H4_0$: Perceived Ease of Use (PEOU) will not be determined by understandability efficiency. $H4_1 = \neg H4_0$. The rationale for these hypotheses is that understandability efficiency represents a performance-based measure of actual efficiency, while PEOU represents a perception-based measure of efficiency. According to the MEM, understandability efficiency measures the effort required to apply the method, which should determine perceptions of effort required.
- $H5_0$: Perceived Usefulness (PU) will not be determined by understandability effectiveness. $H5_1 = \neg H5_{0.}$ The rationale for these hypotheses is that understandability effectiveness represents a performance-based measure of effectiveness, while PU represents a perception-based measure of effectiveness. According to the MEM, perceptions of effectiveness should be determined by actual effectiveness.
- $H6_0$: Perceived Usefulness (PU) will not be determined by Perceived Ease of Use (PEOU). $H6_1 = \neg H6_0$. These hypotheses follow on from the Technology Acceptance Model, in which a direct influence of perceived ease of use on perceived usefulness was found.
- $H7_0$: Intention to Use will not be determined by Perceived Usefulness. $H7_1 = \neg H7_{0.}$ These hypotheses follow on from the Technology Acceptance Model, in which a direct influence of perceived usefulness on intention to use was found.

- $H8_0$: Intention to Use will not be determined by Perceived Ease of Use. $H8_1=\neg H8_0$. These hypotheses follow on from the Technology Acceptance Model, in which a direct influence of perceived ease of use on intention to use was found.

The evaluation model consequently denotes that requirements modeling methods will be adopted in practice on the basis of perceptions of their ease of use and usefulness. We are aware that it would be important to examine and measure the correlation between intention to use and actual usage when employing models based on TAM. However, the actual usage measures actual impact in practice (as opposed to potential impact, defined by Intention to Use). It may be evaluated by surveys of practice (it cannot be used in the evaluation of newly-proposed methods but only in assessing established ones). Typical measures of self-reported usage include frequency of use and intensity of use.

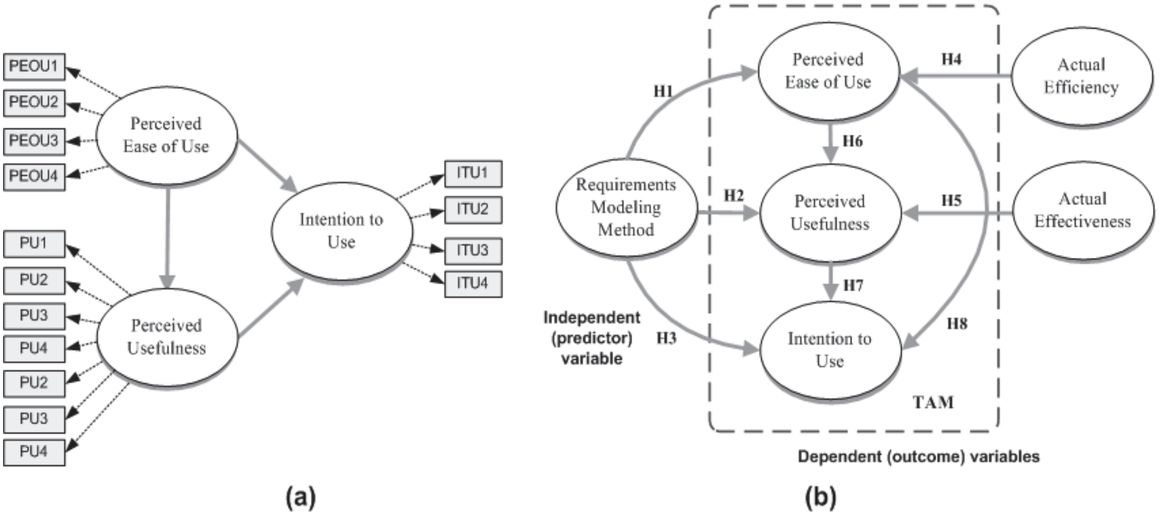


Figure 2. Adapted MEM (a) shows the survey instrument and (b) shows the theoretical model

Table 1 shows the items defined to measure the perception-based variables. The items defined for each construct were combined in a survey, consisting of 15 questions. The items were formulated by using a 5-point Likert scale, using the opposing-statement question format. Various items within the same construct group were randomized, to prevent systemic response bias. PEOU is measured by using 4 items in the survey. PU is measured by using 7 items in the survey. Finally, ITU is measured by using 4 items in the survey. The survey also included three open questions related to the improvement of the requirements modeling method.

Moreover, in order to ensure the balance of items in the survey, half the questions on the left-hand side were negated, to avoid monotonous responses. The resulting measurement instrument can be downloaded at http://www.dsic.upv.es/~sabrahao/REmethodseval/survey.pdf.

Table 1. Items in the survey for measuring the perception-based variables

Item	Item Statement
PEOU1	The requirements modeling method is simple and easy to follow.
PEOU2	Overall, the requirements models obtained by the method was easy to use
PEOU3	It was easy for me to understand what the requirements model was trying to model.
PEOU4	The requirements modeling method is easy to learn.
PU1	I believe this requirements modeling method would reduce the time required to understand software requirements.
PU2	Overall, I found the requirements modeling method to be useful.
PU3	I believe this requirements modeling method is useful for building a conceptual model of a software system.
PU4	I believe that the requirements specifications obtained with this method are organized, clear, concise and non-ambiguous.
PU5	I believe this requirements modeling method has enough expressiveness to represent functional requirements.
PU6	Overall, I think this requirements modeling method provides an effective means of describing requirements specifications.
PU7	Using this requirements modeling method would improve my performance in describing requirements specifications.
ITU1	If I am working in a company in the future, I would use this requirements modeling method to specify functional requirements.
ITU2	It would be easy for me to become skilful in using this requirements modeling method.
ITU3	I intend to use this requirements modeling method in future.
ITU4	I would recommend the use of this requirements modeling method.

The eight hypotheses proposed through the use of the arrows between the variables shown in the theoretical model (H1 to H8 in Figure 2 (b)) were empirically tested through a family of experiments (see Section 5), which was conducted to evaluate the likelihood of acceptance in practice of a RUP extension for requirements modeling, as presented in the following section.

4. RUP EXTENSION FOR REQUIREMENTS MODELING

RUP is a software engineering process that provides a disciplined approach for the definition of tasks and responsibilities in the development of software systems [22][24]. Its goal is to ensure the production of software that satisfies clients' needs by following a use case-driven approach.

Our RUP extension for requirements modeling provides specific techniques for specifying functional requirements. It also defines model transformations for decomposing high-level software requirements into UML design models systematically. The result is a Requirements Model [19] that extends the original RUP disciplines by including tasks and artifacts for the identification, organization, specification, and transformation of requirements, thus enabling a model-driven development approach. In particular, the RUP extension deals with specific tasks and artifacts in the *Requirements* and *Analysis and Design* disciplines. Only those changes considered to be most relevant to the *RUP for small projects* are described, due to limited space [22]. An engineering process architecture called UMA (Unified Method Architecture) is employed for this extension. UMA is based on the OMG (Object Management Group) SPEM 1.1 pattern (Software Process Engineering Meta-Model) for process engineering. The pattern defines schema and terminology with which to represent methods, and consists of method content and processes [17].

During the *Requirements* discipline, the *System Analyst* performs the requirements elicitation by outlining the system's functionality and by delimiting the system. This is done through a number of tasks: developing the vision, finding actors and use cases, structuring the use case model, etc. (see Figure 3). The main extension to this discipline occurs in the redefined task *Structure the Use-Case Model* (shown in Figure 3 with a dotted circle). It is in this task that the use-case model is structured so as to make the requirements easier to understand and to maintain. More specifically, when performing this task we create a new artifact (*Functions Refinement Tree* – FRT, shown in Figure 3 with a solid circle).

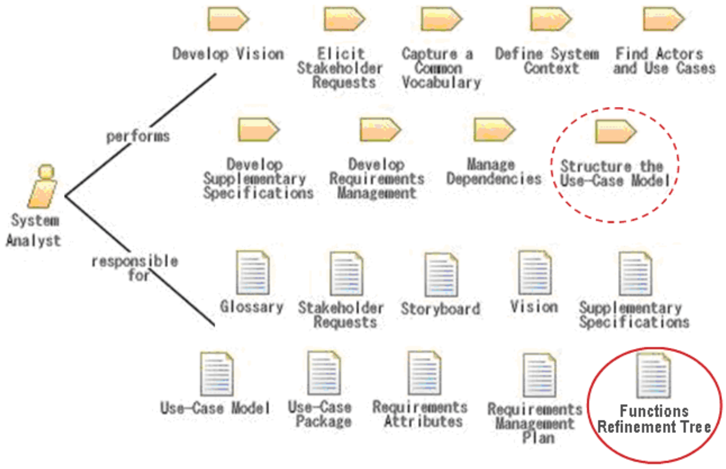

Develop Vision
Elicit Stakeholder Requests
Capture a Common Vocabulary
Define System Context
Find Actors and Use Cases
performs
Develop Supplementary Specifications
Develop Requirements Management
Manage Dependencies
Structure the Use-Case Model
System Analyst
responsible for
Glossary
Stakeholder Requests
Storyboard
Vision
Supplementary Specifications
Use-Case Model
Use-Case Package
Requirements Attributes
Requirements Management Plan
Functions Refinement Tree

Figure 3. Tasks and artifacts of the Requirements discipline

The FRT is a means that is used to identify and organize system functionality. When using this technique, the root is the *Mission Statement* (identified when performing the *Develop Vision* task), the internal nodes are functional groups, and the leaves are the elementary functions of the desired system that correspond to the Use Case concept. A function is regarded as elementary if it is triggered by an event sent by a system user (actor) or by the occurrence of a temporal event.

During the *Analysis and Design* discipline, the *Designer* leads the design of the system within the constraints of the requirements, (which is an extremely important issue to take into account), thus establishing the traceability between requirements and analysis and design. This is done by carrying out a number of tasks: class design, subsystem design, use case analysis, use case design, etc. (see Figure 4).

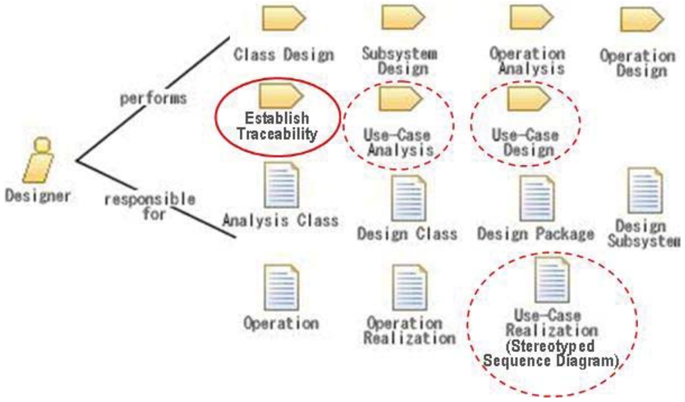


Figure 4. Tasks and artifacts of the Analysis and Design discipline

The main extensions in this discipline occur in the redefined tasks *Use-Case Analysis* and *Use-Case Design* (shown in Figure 4 with dotted circles). The Use-Case Analysis task enables us to perform the transition between Requirements to Analysis and Design, using collaborating objects (interaction diagrams) as a bridge. It therefore provides a mechanism with which to trace behavior in the Analysis and Design Models back to the Use-Case Model, while organizing collaborations around the Use Case concept.

In specific terms, during this task we use Sequence Diagrams to show the required object interactions, using three types of objects: *boundary objects*, to represent the border between the system and the actors, *actor* objects, and *entity* objects. Control objects are not represented at this level, since our concern is to identify *which* object interactions occur, rather than *how* these interactions occur.

The redefined *Use-Case Design* task is very important in our RUP extension. This task defines how to refine the products of *Use-Case Analysis* by developing design-level Use-Case realizations. To be specific, in this task we extend the classical Use Case realizations (Sequence Diagrams) with stereotyped messages, in order to classify the different types of object interactions (messages): *services*, *query*, *signal*, and *connect*. Despite the fact that this information could increase the complexity of object interaction specifications, it is in fact crucial in providing mechanisms that analyze these diagrams automatically and transform them into other analysis and design artifacts.

A further important extension in the *Analysis and Design* discipline is the new task *Establish Traceability* (shown in Figure 4 with a solid circle). This task establishes new traceability links, by analyzing the information contained in the *Use Case Realization artifact* (Stereotyped Sequence Diagram). This analysis is performed by using a Transformation Rules Catalogue [20], which defines a set of model-to-model (M2M) transformation rules. Figure 5 depicts the model-driven requirements approach, including the modified *artifacts built* by following our RUP extension.

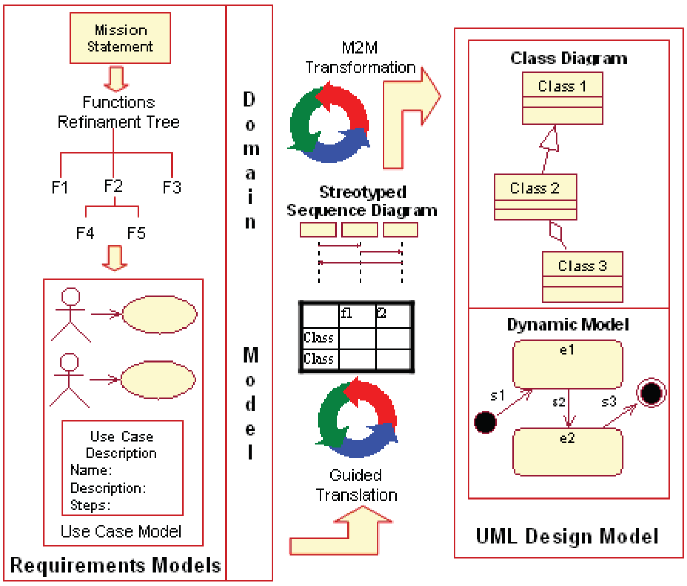


Figure 5. Overview of the Requirements Modeling approach

5. OVERVIEW OF THE FAMILY OF EXPERIMENTS

This section describes the empirical testing of the proposed evaluation method. This was done by using the evaluation method as the theoretical basis for designing a family of experiments to evaluate a specific requirements *modeling* method (i.e., the RUP extension for requirements *modeling).*

A family of experiments contains multiple similar empirical studies that pursue the same goal. As Basili *et al.* [5] remark, a family of experiments builds the knowledge that is needed to extract significant conclusions that can be applied in practice. We performed the family of experiments based on the experimental method proposed by Ciolkowski *et al.* [10] and guided by the experimental process of Wohlin *et al.* [41].

5.1 Experiment preparation

The general goal of the experiments is to test the usefulness of our evaluation model empirically. To this end, our evaluation model was applied to predict the *likelihood of acceptance* of a specific *requirements modeling* method in practice (i.e., the RUP extension shown in Section 4). The Goal-Question-Metric (GQM) template [6] for goal definition was used to define the experimentation goal as follows:

> **Analyze** *requirements modeling artifacts* **for the purpose of** *testing the proposed evaluation method* **with respect to** *its usefulness for predicting the likelihood of acceptance of a specific requirements modeling method in practice (i.e., the RUP extension shown in Section 4)* **from the point of view of the** *researcher*, **in the context of** *a group of undergraduate students and practitioners from two Spanish universities.*

This experimental goal will also allow us to evaluate the reliability and validity of the measurement instrument, as well as the predictive and exploratory power of the MEM. This is achieved by applying the adapted MEM in practice, to test the users' ability to understand the requirements *modeling artifacts* produced when using a particular method for requirements *modeling* (i.e., the RUP extension).

The research questions addressed by the experimentation are:

- RQ1: Is the RUP extension perceived as both easy to use and useful? If so, are the users' perceptions a result of their actual performance in understanding the requirements *modeling artifacts obtained* with the RUP extension?
- RQ2: Is there an intention to use the RUP *extension in the future? If so, is the intention to use it a result of the perceptions experienced by the subjects on understanding the requirements modeling artifacts obtained* with the RUP extension?

These research questions were evaluated by testing a number of hypotheses. As there is currently no standard requirements method for specifying requirements, we cannot evaluate the RUP extension by comparing it to a control method. Hence, the independent variable has only one value in a nominal scale. We must stress, however, that the subjects have an adequate level of experience with UML and RUP. The dependent variables based on the MEM variables (see Figure 2 (b)) were originally defined in section 3.2 and are summarized in Table 2.

Table 2. Summary of the dependent variables

Type	Name	Measure	Scale
Performance-based variables	Understandability effectiveness	$\sum_{i=1}^{n} \frac{CorrectAnswers / NumberOfQuestions}{n}$	Ratio
	Understandability efficiency	$\frac{\sum_{i=1}^{n} (UnderstandabilityEffectiveness_i / Time_i)}{n}$	Ratio
Perception-based variables	Perceived ease of use (PEOU)	Calculated as the mean of the items PEOU1..PEOU4 obtained from the survey shown in Table 1	Ratio
	Perceived usefulness (PU)	Calculated as the mean of the items PU1..PU7 obtained from the survey shown in Table 1	Ratio
	Intention to use (ITU)	Calculated as the mean of the items ITU1..ITU4 obtained from the survey shown in Table 1	Ratio

The first research question (RQ1) was addressed by defining the following hypotheses:

- $H1_0$: The RUP extension is perceived as difficult to use. $H1_1 = \neg H1_0$
- $H2_0$: The RUP extension is perceived as not useful. $H2_1 = \neg H2_0$
- $H4_0$: Perceived ease of use will not be determined by understandability efficiency. $H4_1 = \neg H4_0$

- $H5_0$: Perceived usefulness will not be determined by understandability effectiveness. $H5_1 = \neg H5_0$

The first two hypotheses relate to a direct relationship between the use of the RUP extension and the users' performance and perceptions, while the last two hypotheses were formulated to test the causal relationships between dependent variables (such as performance having an effect on perceptions).

The second research question (RQ2) was addressed through the formulation of the following hypotheses:

- $H3_0$: There is no intention to use the RUP extension in the future. $H3_1 = \neg H3_0$
- $H6_0$: Perceived usefulness will not be determined by perceived ease of use. $H6_1 = \neg H6_0$
- $H7_0$: Intention to use will not be determined by perceived usefulness. $H7_1 = \neg H7_0$
- $H8_0$: Intention to use will not be determined by perceived ease of use. $H8_1 = \neg H8_0$

The first hypotheses relate to a direct relationship between the use of the RUP extension and the users' intentions (see Figure 2 (b)). Nonetheless, the evaluation model also proposes other relationships that indicate causal links between dependent variables (such as performance having an effect on perceptions, or perceptions influencing intentions). The remaining hypotheses were formulated to test the predictive and explanatory power of the MEM.

5.2 Context definition

To make it easier to generalize the results, the following groups of subjects were identified, thereby establishing the context of each individual experiment:

- Undergraduate students. These are final-year Computer Science students at the Universidad Politécnica de Valencia and the University of Castilla-La-Mancha, and they could be considered as the next generation of professionals [23]. It has been shown that, under certain conditions, there is no great difference between this type of students and professionals [18][5]. We therefore believe that their ability to understand requirements models is comparable to that of typical novice analysts.

- Practitioners. These are software engineering professionals who work at the University of Castilla-La-Mancha. They have between 1 and 8 years of experience with static and dynamic *modeling techniques*, including UML use case and sequence diagrams.

5.3 Experimental tasks and materials

With the exception of the fourth experiment, in which a different design was employed, each experiment included two tasks: an understanding task and a post-task survey.

In the understanding task, the subjects were asked to answer a set of six Yes/No questions corresponding to *requirement modeling artifacts*. This task was used to collect data with which to evaluate the performance-based variables. In the post-task survey task, the subjects were then asked to complete a survey, in order to evaluate the RUP extension. This survey is a measurement instrument which is used to collect data to evaluate the perception-based variables.

The material prepared for the family of experiments was composed of training materials, experimental objects (requirements *modeling artifacts*) and a survey instrument. The following training materials were prepared: a set of instructional slides describing the RUP extension (*requirements modeling notation), the procedure for applying it (requirements modeling method*), and an application example with which to illustrate its application.

The experimental objects consisted of 9 sequence diagrams from 3 different applications (a car rental system, a hotel management system, and a singing contest management system). The *stereotyped sequence diagrams were used as experimental objects because they are the main artifacts produced* by the RUP extension. Each sequence diagram included around 6 to 9 objects and 10 to 13 interaction messages. These application examples were selected due to their being excerpts of real-world applications and also because they represent well-known problem domains.

An example of a stereotyped sequence diagram is shown in Appendix A. Each diagram has a corresponding test (Test 1) consisting of 6 Yes/No questions to test the subjects' understanding of the model. The subjects also had to write down the times at which they started the questionnaires, as well as the time they finished completing them.

This understanding task allowed us to obtain two measures for understandability (see Table 2): Understandability *efficiency* (the average of the efficiency obtained in each of the diagrams used), and Understandability *effectiveness* (the average of the effectiveness obtained in each of the diagrams used).

In the post-task survey task, the subjects were asked to fill in the measurement instrument shown in section 3.1. This includes 15 closed questions, which were based on the items used to measure the constructs of the MEM [28].

As the diffusion of the experimental data is important to the external replication of the experiments, we have uploaded all the material used in this family of experiments onto a web site (www.dsic.upv.es/~sabrahao/MEM-Req). This corresponds to the presentation and package activity suggested in the experimental process provided by Wohlin *et al.* [41].

5.4 Conduct individual experiments

Following the general plan for the family of experiments, we carried out four individual experiments, as is shown in Figure 6. The individual experiments were grouped in two main categories, depending on the type of subjects and the experimental design employed:

- Group 1. Undergraduate students from the Universidad Politécnica de Valencia (UPV) and University of Castilla-La-Mancha (UCLM).
- Group 2. Practitioners from the UCLM.

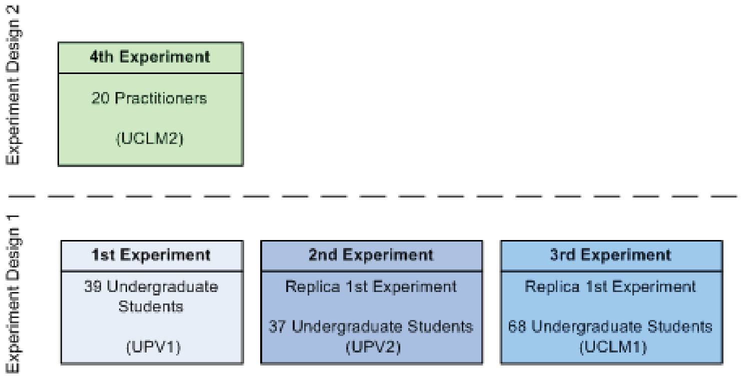


Figure 6. Experiments in the family

No threshold value exists in empirical software engineering as yet with regard to the recommended number of replications needed to test an experimental hypothesis. We carried out three replications of the original experiment. The purpose was to provide evidence for the generalization of a result, by repeating the first experiment in different environments, with different subjects.

In particular, the second and third experiments are *strict* replications of the first experiment (i.e., they duplicate the conditions of the original experiment as accurately as possible), while the fourth experiment is a *differentiated* replication (i.e., experiments that introduce variations into essential aspects of the experimental conditions) [44]. More specifically, the purpose of the fourth experiment was to repeat the first experiment with different subjects and materials (using a different experimental design). Differentiated replications can be conducted to identify potentially important environmental factors that affect the experimental results. In addition, the second experiment is an *internal* replication (i.e., an experiment replicated by the same experimenters to verify whether the results were a one-off, chance occurrence), while the third and fourth experiments are *external* replications of the original experiment (i.e., an experiment replicated by an independent group of experimenters). It is important to note that the people carrying out the replications were unaware of the results of the original experiment.

Each individual experiment was carried out by considering both the general plan established in the context of the family and the feedback obtained as a result. The design of the experiments is described in Section 6.

5.5 Family data analysis

When the data concerning the results of the individual experiments was collected and analyzed, it was not only important to obtain local conclusions (related to each individual experiment); it was also essential to extract the overall conclusions obtained, by carrying out a global analysis of the whole family of experiments. The analysis and interpretation of individual experiments are presented in Section 7, and the data analysis of the family is shown in Section 8.

6. DESIGN OF INDIVIDUAL EXPERIMENTS

In this section we describe the main characteristics of each of the four experiments that constitute our family of experiments.

6.1. The first experiment (UPV1)

6.1.1 Planning

This section details the experimental plan, describing the context, the variables, hypotheses and the experiment design.

Subjects selection. The participants in the first experiment were 39 fourth-year Computer Science students from the Universidad Politécnica de Valencia, who were enrolled on the second Software Engineering course from Sept. 2007 to Jan. 2008. We took a "convenience sample" (i.e., all the students available in the class) [47]. The subjects had 6 months of experience in *modeling* with UML/RUP and 3 years of experience in the Object-Oriented paradigm. In order to avoid persistence effects, the experiment was carried out by subjects who had never done similar experiments. The subjects were encouraged to participate by offering them an extra point in the final mark for performing the required tasks. The students worked on projects involving a variety of software applications such as insurance, banking and hotel reservations. They were given opportunities during the semester to improve their *modeling activity* as part of a project that was developed throughout the semester. We therefore believe that their ability to understand models is comparable to that of a typical novice analyst.

Variables selection. As was mentioned previously, since there is currently no standard requirements method for specifying requirements, we cannot evaluate the RUP extension against a control method. Hence, the independent variable has only one value in a nominal scale. Having said that, we again stress that the subjects had an adequate level of experience with UML and RUP.

The dependent variables based on the MEM variables (see Figure 2 (b)) were originally defined in section 3.2 and are summarized in Table 2. We focus on the understandability of the *requirements modeling artifacts,* since this is essential for the validation of requirements between developers and stakeholders. However, we specifically focus on the understandability of the sequence diagrams produced using the RUP extension since: (a) they capture the main object interactions needed to realize each use case identified. It is in this phase that the analyst builds up the threads that weave the object classes together; (b) they facilitate the transformation of requirements into UML design models.

Hypothesis formulation. The eight hypotheses tested in this experiment were defined in Section 5.1. To make these easier to read, the relations between the variables expressed in the aforementioned hypotheses are illustrated in Table 3.

Table 3. Summary of hypotheses

Hypotheses	Relation between variables
H1	RUP extension→Perceived Ease of Use (PEOU)
H2	RUP extension→Perceived Usefulness (PU)
H3	Understandability Efficiency→Perceived Ease of Use (PEOU)
H4	Understandability Effectiveness→Perceived Usefulness (PU)
H5	RUP extension→Intention to Use (ITU)
H6	Perceived Ease of Use (PEOU)→Perceived Usefulness (PU)
H7	Perceived Usefulness (PU)→Intention to Use (ITU)
H8	Perceived Ease of Use (PEOU)→Intention to Use (ITU)

Instrumentation. The experimental objects consisted of 9 sequence diagrams from 3 different case studies (car rental, hotel management, and singing contest management systems).

6.1.2 Operation

The operation was performed through the following activities:

Preparation. The subjects were given a training session before the experiment took place. However, they were not aware of what aspects we intended to study. The training session consisted of two hours. In the first hour, we explained the main concepts of the RUP extension (e.g., *the modeling notation* for sequence diagrams) and demonstrated their application with examples. In the remaining hour, the subjects used the RUP extension to solve an exercise similar to the kind that would be tackled during the execution of the experiment.

Execution. The experiment took place in a single room. It was controlled in such a way that no interaction between subjects occurred. Each subject was assigned all the materials (see Section 5.3), with the *9 tests (balanced within subjects). The requirements modeling artifacts were assigned* in a different order, to limit learning effects. The subjects were shown how to carry out the tasks. After they had finished the understanding task, the subjects were asked to perform the post-task survey. To avoid a possible ceiling effect, there was no real time limit for the performance of the tasks. Moreover, in order to prevent a potential bias in subject responses (i.e., the risk of wishing to please the experimenters by *favorable* judgments of the RUP extension), the subjects were told that their answers would be treated anonymously. Before filling in the survey, the students were also informed that their grade on the course would not be affected by their performance in the experiment.

Data recording and validation. The performance-based dependent variables were measured by using a data collection form that was included in the replication package, available at (www.dsic.upv.es/~sabrahao/MEM-Req). After the experiment took place, we collected the experimental data, which consisted of a table of 351 rows (9 diagrams x 39 subjects) and 5 columns (Understandability Effectiveness, Understandability Efficiency, PEOU, PU, and ITU). We then performed "data cleaning", in order to exclude any observations that were not complete because the subjects had not written down the time or because they had not answered the survey. All the questions were answered in each questionnaire, thus assuring the completeness of the tasks performed. Hence, the final data for testing the hypotheses were 325 observations. Preliminary results of this experiment can be found in [2].

6.2. The second experiment (UPV2)

In order to confirm the results obtained in the first experiment, we replicated this experiment under the same conditions (strict replication), changing only the subjects [5]. Strict replications are needed to increase confidence in the conclusion validity of the experiment.

The subjects were a different group of 37 students from the same course as those of the original experiment (Software Engineering II at the Universidad Politécnica de Valencia, September 2007 to January 2008). As in the first experiment, this experiment was organized as a compulsory part of the course. Exactly the same materials and experimental objects were used.

The second experiment was performed on the same day as the original one, and immediately after the first experimental run. We do not therefore believe that having subjects who are students from the same course has influenced the replica results, as there was no time for the students from both groups to communicate with each other.

6.3. The third experiment (UCLM1)

The third experiment was also an external replication of the first experiment (strict replication). The subjects were 68 undergraduate students, aged between 22 and 24, enrolled on a Software Engineering course at the University of Castilla-La-Mancha (UCLM) during 2008. The same experimental objects as those used in the first two experiments were employed.

6.4. The fourth experiment (UCLM2)

This experiment was an external replication which varied the manner in which the other experiments were run. The purpose was to increase the confidence in the experimental results by testing the same hypotheses as those previously mentioned, but altering certain characteristics of the original experiment, in order to increase external validity; details of this will be set out in the following sub-sections.

6.4.1 Planning

The planning stage was composed of the following activities:

Subjects selection. The context of the replica was a group of practitioners who were at the University of Castilla-La-Mancha (UCLM) in April 2008. They were people with work contracts, involved in several industrial transfer projects at the university. The participants were between 24 and 36 years of age and they had different backgrounds. They had *between 1 and 10 years of working experience, with 1 to 8 years of experience in modeling with UML* and RUP. However, their level of training in the RUP extension was the same as that of the undergraduate students who had done the previous experiment in the family.

Variables selection. The same variables as those used in the other experiments were selected. Moreover, taking into account certain suggestions concerning the measurement of understandability [7][16], we have changed the experimental design to consider two new variables (from the Cognitive Theory of Multimedia Learning [27]), in order to measure other dimensions of model understandability:

- *Retention:* this measures the comprehension of the material being presented, and the ability to retain knowledge from it.
- *Transfer:* this measures the ability to use the knowledge gained from the material to solve related problems which are not directly answerable from the material.

The choice to use the above theory was made for several reasons. Firstly, it focuses on words and graphics, which are the elements in the UML sequence diagram notation. Secondly, it provides principles for the design of effective multimedia presentations which can be tested empirically. Thirdly, it has evolved through years of work and development of experimental instruments and methods related to model comprehension [27][26].

Hypothesis formulation. The eight hypotheses defined in Section 5.1 were tested. We also formulated other hypotheses to test the influence of the new variables (retention and transfer) on the user perceptions of the quality of the *requirements modeling artifacts* of the RUP extension:

- $H9_0$: Increases in the Retention Efficiency will not cause increases in Perceived Ease of Use (PEOU). $H9_1=\neg H9_0$
- $H10_0$: Increases in the Transfer Efficiency will not cause increases in Perceived Ease of Use (PEOU). $H10_1=\neg H10_0$
- $H11_0$: Increases in the Retention Effectiveness will not cause increases in Perceived Usefulness (PU). $H11_1=\neg H11_0$

- $H12_0$: Increases in the Transfer Effectiveness will not cause increases in Perceived Usefulness (PU). $H12_1=\neg H12_0$

These new variables were added as an *operationalization of both* the ‘actual efficiency’ and the ‘actual effectiveness’ constructs of the MEM, since we measured both the effort and the correctness of the subject responses in their completion of tests 2 and 3 (described below).

Instrumentation. The experimental objects consisted of 6 sequence diagrams from 3 different case studies (car rental, hotel management, and singing contest management systems). Each sequence diagram had three tests:

- Test 1: this was composed of understandability tasks (the same as those used in previous experiments).
- Test 2: this consisted of a ‘fill-in-the-blanks’ task, in which the subjects had to complete a test describing the functionality of the diagrams (see Appendix A for an example). This activity was used to calculate the efficiency and effectiveness of the retention task.
- Test 3: the subjects had to name a set of new messages that had been attached to the original version of the diagrams, but which were *labeled* only with the parameters (see Appendix A for an example). This task was used to calculate the efficiency and effectiveness of the transfer task.

Each subject received the 6 requirements *modeling artifacts, including* the 3 tests, in a different order.

6.4.2 Operation

We began with a pre-test to measure the subjects’ level of knowledge and familiarity with UML *modeling and* RUP. We then applied a sequence of tasks in a particular order to ensure the internal validity of the replication.

The subjects first completed the multiple choice understandability tasks (Test 1) with the sequence diagram provided. This ensured that the subjects scanned the whole model so that they would be ready to do the next tasks. After the understandability task, the models were taken away. The subjects then completed the retention task, after which they carried out the transfer task. The aim of eliminating the models is important, in that it ensured that the only information available to the subjects was the cognitive model produced when they viewed the model.

7. ANALYSIS AND INTERPRETATION

We present the validation of the measurement instrument first, followed by the analysis of the results for each individual experiment, according to the research questions posed. All the results were obtained by using SPSS 13.0 for Windows. The collected data were analyzed according to the following levels of significance: not significant (p>0.1), low significance (p<0.1), medium significance (p<0.05), high significance (p< 0.01) and very high significance (p< 0.001).

7.1 Assessing reliability and validity of the measurement instrument

The construct validity of the survey instrument with the PEOU, PU and ITU items was evaluated by using an inter-item correlation analysis [8]. The evaluation was based on two criteria: convergent and discriminant validity.

Convergent validity refers to the convergence among the different items that are meant to measure the same construct. The correlation between such items should be high. The convergent validity value of an item is calculated as the average correlation between the scores for that item and the scores for the other items that are intended to measure the same construct.

Discriminant validity refers to the divergence of the items used to measure different constructs. The correlation between items used to measure different constructs should be low. The discriminant validity value of an item is calculated as the average correlation between the scores for that item and the scores for the items that are intended to measure another construct. Low divergent validity values indicate high discriminant validity. According to Campbell and Fiske [8] an item's convergent validity value must be higher than its divergent validity value; otherwise the scores on the item should not be used in the data analysis.

The results of the inter-item correlation analysis (see Table B1 of the Appendix B for an example) for the 4 data sets were:

- UPV1: Q7 (a PU item) did not pass the Campbell and Fiske test (i.e., divergent validity was greater than convergent validity). Q7 was therefore excluded from the analysis. The average convergent validity was 0.39 for PEOU, 0.32 for PU, and 0.44 for ITU.
- UPV2: again, Q7 did not pass the validity test. With the exclusion of this item, the average convergent validity was 0.31 for PEOU, 0.36 for PU, and 0.50 for ITU.
- UCLM1: Q15 did not pass the validity test. With the exclusion of this item, the average convergent validity was 0.34 for PEOU, o.38 for PU, and 0.40 for ITU.
- UCLM2: Q7 and Q12 did not pass the validity test. With the exclusion of these items, the average convergent validity was 0.35 for PEOU, o.35 for PU, and 0.53 for ITU.

The use of multiple items to measure a same construct also requires the examination of the reliability or internal consistency among the measurements. We evaluated the reliability of the survey by using Cronbach's alpha. For this analysis, the items that did not pass the validity test were excluded from their corresponding data sets. The results of the reliability analysis are as follows: UPV1 (PEOU = 0.73; PU = 0.75; ITU = 0.73), UPV2 (PEOU = 0.66; PU = 0.80; ITU = 0.78), UCLM1 (PEOU = 0.70; PU = 0.806; ITU = 0.70), and UCLM2 (PEOU = 0.73; PU = 0.5; ITU = 0.78).

Almost all the constructs have an alpha value equal to, or greater than 0.7, which is a common reliability threshold [30]. As a result of this analysis, we can conclude that the items in the survey are reliable and valid measures of the underlying perception/intention-based constructs of the proposed evaluation model.

7.2 Analysis of user perceptions and likelihood of acceptance

Table 4 shows descriptive statistics for the performance-based variables. In the first experiment, the results indicate that, as regards time taken to perform the understandability tasks (efficiency), the subjects range from 2.40 to 10.10 minutes. In terms of understandability effectiveness, the percentage of correctness ranges from 73.80% to 97.20%. The replications show similar results.

In terms of retention, the practitioners took 8.84 to 18.43 minutes to complete the fill-in-the blank tasks, with a percentage of correctness that varied from 82.22% to 98.89%, which is a very good level of retention. Finally, the knowledge transfer tasks were completed within a range of 1.13 to 3.72 minutes, with a percentage of correctness of 33.30% to 92.20%.

Table 4. Descriptive statistics for the performance-based variables

Data set	Variable	Min.	Max.	Mean	Std. Dev.
UPV1	Understandability Efficiency (min.)	2.40	10.10	5.43	1.733
	Understandability Effectiveness (%)	73.80	97.20	86.92	4.725
UPV2	Understandability Efficiency (min.)	3.68	8.70	5.29	1.286
	Understandability Effectiveness (%)	76.67	95.83	87.46	4.766
UCLM1	Understandability Efficiency (min.)	3.00	8.70	5.17	1.299
	Understandability Effectiveness (%)	61.90	89.58	77.23	7.157
UCLM2	Understandability Efficiency (min.)	2.90	7.00	4.528	1.239
	Understandability Effectiveness (%)	63.89	97.44	82.62	8.970
	Retention Efficiency (min.)	8.84	18.43	13.09	2.893
	Retention Effectiveness (%)	82.22	98.89	92.99	4.391
	Transfer Efficiency (min.)	1.13	3.72	2.410	0.7345
	Transfer Effectiveness (%)	33.30	92.20	69.68	14.008

For the perception-based variables (PEOU, PU, ITU) the results of the experiment and its replications showed that the subjects perceived the RUP extension as easy to use and useful. They also expressed an intention to use it in the future. This can be observed by the mean value for these variables (i.e., all of them are greater than the neutral score (i.e., the score 3).

In order to evaluate the likelihood of acceptance in practice of the RUP extension, we tested hypotheses H1, H2 and H3 (see Figure 2 (b)). These hypotheses were tested by verifying whether the scores that the subjects assign to the constructs of the MEM are significantly better than the neutral score on the Likert scale for an item. The scores of a subject are averaged over the items that are relevant for a construct. We thus obtained three scores for each subject.

The Kolmogorov-Smirnov test for normality was applied to the PEOU, PU and ITU data. As the distribution was normal, we used the One-tailed sample t-test to check for a difference in mean PEOU, PU, and ITU for the RUP extension and the value 3. The results shown in Table 5 allow us to reject the null hypotheses, meaning that we corroborated empirically that the subjects perceived the RUP extension as being easy to use and useful, and that there is an intention to use it in the future. The statistical sig. for the hypotheses was very high ($p<0.001$).

Table 5. Descriptive statistics and 1-tailed one sample t-test rank for perception-based variables

Data set	Variable	Min.	Max.	Mean	Std. Dev.	Std. error mean	t	1-tailed p-value
UPV1	PEOU	2.25	4.75	3.69	0.650	0.6973	6.611	p<0.001
	PU	2.33	4.67	3.77	0.510	0.7719	9.325	p<0.001
	ITU	2.25	4.75	3.63	0.619	0.6315	6.281	p<0.001
UPV2	PEOU	2.75	2.75	3.78	0.462	0.7857	10.054	p<0.001
	PU	2.33	4.83	4.00	0.536	1.0095	11.139	p<0.001
	ITU	1.50	4.75	3.96	0.624	0.96429	9.132	p<0.001
UCLM1	PEOU	1.25	5.00	3.38	0.691	0.0838	4.602	p<0.001
	PU	1.57	4.86	3.53	0.636	0.0771	6.887	p<0.001
	ITU	1.33	5.00	3.21	0.789	0.0957	2.254	p=0.013
UCLM2	PEOU	2.25	4.75	3.80	0.684	0.1439	5.422	p<0.001
	PU	2.20	4.80	3.87	0.567	0.1238	7.077	p<0.001
	ITU	1.50	4.75	3.64	0.735	0.1605	4.005	p<0.001

7.3 Hypothesis testing

The aim of this section is to validate the structural part of the MEM in terms of the causal relationships between its constructs, with the exception of Actual Usage. To do this, we chose regression analysis, since the hypotheses to be tested are causal relationships between continuous variables.

7.3.1 Actual Efficiency vs. Perceived Ease of Use

We tested hypothesis H4 to verify whether perceptions of efficiency are determined by actual efficiency. A simple regression model was built for each data set (the experiment and the replications). Understandability efficiency was used as the independent variable and perceived ease of use as the dependent variable. The regression equations resulting from the analysis are:

$$\textit{Perceived Ease of Use}_{UPV1} = 2.50 + 0.22 * \textit{Understandability Efficiency} \quad (3)$$

$$\textit{Perceived Ease of Use}_{UPV2} = 2.663 + 0.21 * \textit{Understandability Efficiency} \quad (4)$$

$$\textit{Perceived Ease of Use}_{UCLM1} = 1.92 + 0.28 * \textit{Understandability Efficiency} \quad (5)$$

$$\textit{Perceived Ease of Use}_{UCLM2} = 2.52 + 0.28 * \textit{Understandability Efficiency} \quad (6)$$

The regression was found to be very highly significant, with p<0.001 in all data sets, except for UCLM2, in which the significance was medium (see Table 6). In the UPV1 data set, the r2 statistic shows that the Understandability Efficiency accounts for 34% of the variance in perceived ease of use, whereas the replications the Understandability Efficiency explain 35% (UPV2), 28% (UCLM1) and 27% (UCLM2) of the variance in perceived ease of use. The regression coefficient (b) defines the effect of the independent variable on the dependent variable. In this case, an increase in efficiency of one minute results in an increase of 0.21 to 0.28 in PEOU. The Pearson correlation coefficient (r) indicates that there is a good correlation between the variables. This signifies that H4 was confirmed in all data sets.

The UCLM2 data set also includes two new variables (retention and transfer) for measuring other dimensions of actual efficiency. The models for these variables are as follows:

$$\textit{Perceived Ease of Use}_{UCLM2} = 2.41 + 0.11 * \textit{Retention Efficiency} \qquad (7)$$

$$\textit{Perceived Ease of Use}_{UCLM2} = 3.06 + 0.31 * \textit{Transfer Efficiency} \qquad (8)$$

According to Table 6, the regression for Retention Efficiency was found to be highly significant, whereas the regression for Transfer Efficiency was of low significance. According to Mayer [26], these results suggest that fragmented learning occurred, since the retention is high but the transfer is low. Despite these results, the r2 statistic shows that the Retention Efficiency and Transfer Efficiency variables account for 21% and 11% of the variance in perceived ease of use, respectively. This signifies that H9 and H10 were confirmed.

Table 6. Simple regression between Perceived Ease of Use and Actual Efficiency (α = 0.05)

Data set	Regression model	Unstd. coef. (b)	Std. Error	Std. coef. (beta)	T	Sig. (p)	R
UPV1	Constant	2.504	0.289		8.673	p<0.001	
	Understandability Efficiency	0.220	0.051	0.585	4.333	p<0.001	0.585
UPV2	Constant	2.659	0.275		9.683	p<0.001	
	Understandability Efficiency	0.213	0.050	0.592	4.218	p<0.001	0.592
UCLM1	Constant	1.924	0.296		6,502	p<0.001	
	Understandability Efficiency	0.283	0.056	0.531	5.092	p<0.001	0.531
UCLM2	Constant	2.521	0.509		4.956	p<0.001	
	Understandability Efficiency	0.284	0.109	0.515	2.621	p=0.008	0.515
	Constant	2.406	0.648		3.715	p=0.001	
	Retention Efficiency	0.107	0.048	0.453	2.216	p=0.039	0.453
	Constant	3.065	0.507		6.045	p<0.001	
	Transfer Efficiency	0.309	0.202	0.332	1.533	p=0.071	0.332

7.3.2 Actual Effectiveness vs. Perceived Usefulness

We tested hypothesis H5 to verify whether perceived usefulness is determined by actual effectiveness. The resulting regression equations are:

$$\textit{Perceived Usefulness}_{UPV1} = -1.53 + 0.06 * \textit{Understandability Effectiveness} \tag{9}$$

$$\textit{Perceived Usefulness}_{UPV2} = -0.21 + 0.05 * \textit{Understandability Effectiveness} \tag{10}$$

$$\textit{Perceived Usefulness}_{UCLM1} = 0.66 + 0.04 * \textit{Understandability Effectiveness} \tag{11}$$

$$\textit{Perceived Usefulness}_{UCLM2} = 1.43 + 0.03 * \textit{Understandability Effectiveness} \tag{12}$$

For the UPV1 data set, the regression coefficient for effectiveness was very high ($p<0.001$). The r2 statistic shows that effectiveness accounts for 32% of the variance in perceived usefulness, which presents a very high value, given that there could have been other ways to measure effectiveness (e.g. modifiability). In the replications, the regression was found to be highly significant for UPV2, UCLM1 and UCLM2 ($p<0.01$). The r2 statistic shows that effectiveness accounts for 18% (UPV2), 17% (UCLM1) and 22% (UCLM2) of the variance in perceived usefulness (see Table 7). This signifies that H5 was also confirmed in all data sets.

In addition, the UCLM2 data set included two new variables for measuring other dimensions of actual effectiveness. The regression models for these variables are as follows:

$$\textit{Perceived Usefulness}_{UCLM2} = -3.49 + 0.08 * \textit{Retention Effectiveness} \tag{13}$$

$$\textit{Perceived Usefulness}_{UCLM2} = 2.99 + 0.01 * \textit{Transfer Effectiveness} \tag{14}$$

According to Table 7, the regression for Retention Effectiveness was found to be highly significant, whereas the regression for Transfer Effectiveness was low in significance.

Table 7. Simple regression between Perceived Usefulness and Actual Effectiveness ($\alpha = 0.05$)

Data set	Regression model	Unstd. Coef. (b)	Std. Error	Std. coef. (beta)	T	Sig. (p)	R
UPV1	Constant	-1.529	1.293		-1.183	p=0.022	
	Understandability Effectiveness	0.061	0.015	0.565	4.106	p<0.001	0.565
UPV2	Constant	-0.215	1.549		-0.139	p=0.045	
	Understandability Effectiveness	0.048	0.018	0.429	2.731	p=0.005	0.529
UCLM1	constant	0.660	0.771		0.855	p=0.197	
	Understandability Effectiveness	0.037	0.010	0.418	3.740	p<0.001	0.418
UCLM2	constant	1.431	1.065		1.343	p=0.097	
	Understandability Effectiveness	0.030	0.013	0.468	2.308	p=0.016	0.468
	constant	-3.492	2.180		-1.602	p=0.063	
	Retention Effectiveness	0.079	0.023	0.613	3.384	p=0.001	0.613
	Constant	2.985	0.626		4.768	p<0.001	
	Transfer Effectiveness	0.013	0.009	0.316	1.451	p=0.081	0.316

Once more, these results indicate that fragmented learning occurred. This suggests memorization, rather than meaningful learning. The r2 statistic shows that the Retention Effectiveness and Transfer Effectiveness variables account for 38% and 10% of the variance in perceived usefulness, respectively. This means that H11 and H12 were also confirmed.

7.3.3 Perceived Ease of Use vs. Perceived Usefulness

We tested H6 to verify whether Perceived usefulness is determined by perceived ease of use. A simple regression model was built using perceived ease of use as the independent variable and perceived usefulness as the dependent variable. The regression equations are as follows:

$$\textit{Perceived Usefulness}_{UPV1} = 3.32 + 0.\ 12 * \textit{Perceived Ease of Use} \quad (15)$$

$$\textit{Perceived Usefulness}_{UPV2} = 1.72 + 0.61 * \textit{Perceived Ease of Use} \quad (16)$$

$$\textit{Perceived Usefulness}_{UCLM1} = 1.73 + 0.53 * \textit{Perceived Ease of Use} \quad (17)$$

$$\textit{Perceived Usefulness}_{UCLM2} = 3.43 + 0.12 * \textit{Perceived Ease of Use} \quad (18)$$

As is shown in Table 8Table 8 the regression coefficient for the UPV1 data set was found to be not significant (p>0.1). With regard to the predictive power of the model, perceived ease of use explains only 2% of the variance in perceived usefulness as indicated by (r2). The same occurred with the UCLM2 data set.

However, the results obtained in the replications allowed us to corroborate empirically that perceived usefulness is to some extent determined by perceived ease of use. The r2 statistic shows that perceived ease of use accounts for 27% (UPV2) and 34% (UCLM1) of the variance in perceived usefulness. The regression coefficients for perceived ease of use were very high (p<0.001) for these data sets meaning that H6 can be partially confirmed.

Table 8. Simple regression between Perceived Ease of Use and Perceived Usefulness (α = 0.05)

Data set	Regression models	Unstd. coef. (b)	Std. Error	Std. coef. (beta)	T	Sig. (p)	R
UPV1	constant	3.318	0.485		6.845	p<0.001	
	Perceived ease of use	0.123	0.129	0.156	0.949	p=0.174	0.156
UPV2	constant	1.715	0.656		2.613	0.007	
	Perceived ease of use	0.606	0.172	0.523	3.521	p<0.001	0.623
UCLM1	constant	1.727	0.319		5,414	p<0.001	
	Perceived ease of use	0.533	0.092	0.579	5.773	p<0.001	0.579
UCLM2	constant	3.427	0.728		4.705	p<0.001	
	Perceived ease of use	0.118	0.188	0.142	0.627	p=0.269	0.142

7.3.4 Intention to Use vs. Perceived Usefulness

We tested hypotheses H7 to verify whether intention to use is determined by perceived usefulness. The regression equations for the data sets are as follows:

$$\textit{Intention to Use}_{UPV1} = 1.37 + 0.60 * \textit{Perceived Usefulness} \quad (19)$$

$$\textit{Intention to Use}_{UPV2} = 1.04 + 0.73 * \textit{Perceived Usefulness} \quad (20)$$

$$\textit{Intention to Use}_{UCLM1} = 0.42 + 0.79 * \textit{Perceived Usefulness} \quad (21)$$

$$\textit{Intention to Use}_{UCLM2} = 1.43 + 0.57 * \textit{Perceived Usefulness} \quad (22)$$

The regression coefficients were found to be highly significant ($p<0.01$) for UPV1 and UCLM2, and very highly significant for the other data sets. With regard to the predictive power of the model, PU explains 24% (UPV1), 39% (UPV2), 41% (UCLM1) and 19% (UCLM2) of the variance in ITU, as is shown in Table 9. This indicates that perceptions in intention to use are determined by perceptions in usefulness; these results allow us to empirically corroborate that ITU is determined by PU, thereby showing that H8 was confirmed.

Table 9. Simple regression between Perceived Usefulness and Intention to Use ($\alpha = 0.05$)

Data set	Regression model	Unstd. coef. (b)	Std. Error	Std. coef. (beta)	T	Sig. (p)	R
UPV1	constant	1.374	0.670		2.050	p=0.024	
	Perceived usefulness	0.598	0.176	0.493	3.396	p=0.001	0.493
UPV2	constant	1.037	0.639		1.622	p=0.057	
	Perceived usefulness	0.730	0.158	0.627	4.620	p<0.001	0.627
UCLM1	constant	0.419	0.421		0.994	p=0.016	
	Perceived usefulness	0.792	0.117	0.639	6.743	p<0.001	0.639
UCLM2	constant	1.432	1.046		1.370	p=0.035	
	Perceived usefulness	0.570	0.267	0.440	2.135	p=0.023	0.440

7.3.5 Intention to Use vs. Perceived Ease of Use

We tested hypothesis H8 to verify whether intention to use is determined by perceived ease of use. A simple regression model was built for all data sets using perceived ease of use as the independent variable and intention to use as the dependent variable. The regression equations are:

$$\textit{Intention to Use}_{UPV1} = 2.18 + 0.39 * \textit{Perceived Ease of Use} \quad (23)$$

$$Intention\ to\ Use_{UPV2} = 1.53 + 0.64 * Perceived\ Ease\ of\ Use \tag{24}$$

$$Intention\ to\ Use_{UCLM1} = 1.89 + 0.39 * Perceived\ Ease\ of\ Use \tag{25}$$

$$Intention\ to\ Use_{UCLM2} = 2.14 + 0.39 * Perceived\ Ease\ of\ Use \tag{26}$$

The regression coefficient was of medium significance ($p<0.005$) for the UPV1 data set. With regard to the predictive power of the model, perceived ease of use explains 17% of the variance in ITU. This indicates that perceptions in intention to use are, to some degree, determined by perceptions in ease of use. As noted in Table 10, the results of the regression for the replications also allow us to corroborate empirically that intention to use is determined by perceived ease of use. The r2 shows that these two variables account for 23% (UPV2), 12% (UCLM1), and 13% (UCLM2) of the variance in intention to use. We thus believe that sufficient evidence has been found to support H7.

Table 10. Simple regression between Perceived Ease of Use and Intention to Use ($\alpha = 0.05$)

Data set	Regression model	Unstd. coef. (b)	Std. Error	Std. coef. (beta)	T	Sig. (p)	r
UPV1	Constant	2.181	0.543		4.014	p<0.001	
	Perceived ease of use	0.392	0.145	0.412	2.710	p=0.005	0.412
UPV2	Constant	1.532	0.789		1.942	p=0.030	
	Perceived ease of use	0.643	0.207	0.476	3.105	p=0.002	0.476
UCLM1	Constant	1.886	0.456		4.140	p<0.001	
	Perceived ease of use	0.393	0.132	0.344	2.978	p=0.002	0.344
UCLM2	Constant	2.139	0.887		2.411	p=0.013	
	Perceived ease of use	0.395	0.229	0.367	1.721	p=0.051	0.367

8. FAMILY DATA ANALYSIS

This section provides a summary of the results obtained. First of all, we present an analysis of the results in the context of the whole family of experiments. Then we discuss the limitations of the proposed evaluation method, along with the limitations of the empirical test conducted to validate it.

8.1 Summary of results

Once the individual experiments had been carried out, we performed a global analysis of the results in the context of the family of experiments, to determine whether the primary goal of the empirical test had been achieved. We also looked at the various data sets, searching for differences. A summary of the results obtained in each individual experiment is provided in Table 11.

Table 11. Summary of the results of the family of experiments

Data set	Experiments	Nr. of subjects	Type of subjects	Exp. design	Hypotheses confirmed	Hypotheses not confirmed
UPV1	1st experiment	39	Undergraduate students	1	H1, H2, H3, H4, H5, H7, and H8.	H6
UPV2	Strict replication of first experiment (different environment and subjects)	37	Undergraduate students	1	H1, H2, H3, H4, H5, H6, H7, and H8.	none
UCLM1	Strict replication of first experiment (different environment and subjects)	68	Undergraduate students	1	H1, H2, H3, H4, H5, H6, H7, and H8.	none
UCLM2	Replication that varies the manner in which the 1st experiment is run	21	Practitioners	2	H1, H2, H3, H4, H5, H7, H8, H9, H10, H11, and H12	H6

Four experiments were performed, in which 165 subjects belonging to the following groups participated: undergraduate students and practitioners. The following global conclusions were obtained for each research question:

Research question 1: *"Is the RUP extension perceived as easy to use and useful? If so, are the users' perceptions a result of their actual performance in understanding the requirements modeling artifacts obtained with the RUP extension?"*. As expected, the majority of the participants found the RUP extension quite useful and easy to use in performing the specification of functional user requirements. This is supported by their efficiency and effectiveness in performing the understandability tasks. We found support for hypotheses H1 and H2 in all the experiments. This result is encouraging in terms of our intention to use this method as part of a technology transfer effort.

With regard to the users' perceptions, Table 6 and Table 7 show the results of the regression analysis in terms of statistical significance (p-value). As can be noted, H4 and H5 were confirmed in the original experiment and the three *replications. Perceptions of usefulness and ease of use when understanding the requirements modeling artifacts* were therefore influenced by perceptions of subjects' effectiveness and efficiency, respectively. These findings are consistent with the results obtained by Moody [28].

However, the regression results for Retention (efficiency and effectiveness) were found to be highly significant, whereas the regression results for Transfer (efficiency and effectiveness) were of low significance. These results suggest that fragmented learning occurred. The process of applying the RUP extension should therefore be improved to optimize learning.

Research question 2: *"Is there an intention to use the RUP extension in the future? If so, is the intention to use it a result of the perceptions experienced by the subjects on understanding the requirements modeling artifacts obtained with the RUP extension?".* The majority of the respondents were very positive about the use of the RUP extension in the future, which was also supported by the open questions enclosed in the survey instrument. Hypothesis H3 was confirmed in all data sets. This supports the causal relationships put forward by the MEM [13] and the Theory of Reasoned Action (TRA) [14].

With regard to the users' perceptions, the results shown in Tables 8-10 revealed that with the exception of the Perceived Ease of Use vs. Perceived Usefulness relationship (H6), which was not confirmed in the UPV1 and UCLM2 data sets, all the relationships were confirmed. One possible explanation may be that perceived usefulness is determined by other factors other than perceived ease of use. In fact, the UCLM2 data set confirmed this assumption, since the retention effectiveness and the transfer effectiveness variables explain 38% and 10% of the variance in perceived usefulness, respectively.

We *are aware that there may be other factors that could affect people*'*s decisions when using a requirements modeling* method (e.g., tool infrastructure adoption, standardization). However, there are uncontrollable factors (i.e., organizational time frame), partially controllable factors (i.e., systems development backlog) and fully controllable factors (i.e., end-user computing). Our objective here was to select variables that can be controlled, such as the *behavior of* those people using a *requirements modeling method*. We considered perceived ease of use and perceived usefulness, because they are the most important factors in explaining system use [28][13]. The objective was to provide a base for tracing the impact of external variables (understandability) on internal beliefs, attitudes, and intentions. The results thus provide further evidence for the usefulness of MEM in evaluating other types of methods.

In general, the results support the four research questions stated. The results suggest that there were no noteworthy differences between the two types of users (undergraduate students and practitioners) and the two treatments used in the family of experiments. The results of the fourth experiment are in line with those of the first, second and third experiments. Running a family of experiments (including replications), rather than a single experiment provides us with more evidence of the external validity, and thus of the *generalizability, of* the study results. The same hypotheses were tested and confirmed (with few exceptions) in four different environments, using two different experimental objects and two types of participants. Each replication provides further evidence of the confirmation of the hypotheses. We can thus conclude that the general goal of the empirical validation has been achieved.

8.2 Limitations of the study

This section discusses both the limitations of the proposed evaluation method and the limitations of the empirical tests conducted to validate it.

With regard to the proposed evaluation method, two limitations should be acknowledged and addressed in relation to the study presented here. The first limitation concerns the investigation of other factors that may affect the use of *requirements modeling methods in practice. Although our study indicates that the selected performance-based variables influence the likelihood of acceptance of requirements modeling methods in practice*, we should investigate the influence of other variables in future replications of this study. For instance, Davis *et al.* [45] have identified 24 quality attributes of a good software requirement specification (e.g., traceable, verifiable unambiguous, understandable, precise). We plan to use some of these quality attributes as dependent variables for measuring actual effectiveness, since this construct is related to the quality of *the artifact(s)* obtained by applying the method.

The second limitation concerns the measurement scales for measuring the perception-based variables. The only studies that we were able to identify on method adoption were those published by Moody [46]. However, these studies focused specifically on Information Systems design methods. We transferred the TAM items to the context of *requirements modeling methods*. The weakness involved in this is that TAM has been developed specifically for the context of computer usage, and the items may not be totally transferable to a different domain. Although our family of experiments showed good results, an in-depth analysis of this issue should be carried out in further experimentation.

With regard to the empirical testing of the proposed method, it was designed in such a way as to alleviate the following threats to the internal validity:

- Differences among subjects. The tests were randomly assigned to the subjects in a different order. This procedure cancelled out possible learning from similarities in the treatments.
- *Knowledge of the universe of discourse. We used the same requirements modeling artifacts for* all the subjects. This specified the requirements of car rental, hotel management, and singing contest management systems. These are well-known universes of discourse.
- Fatigue effects. On average, the subjects took 5.43 min. to solve the understandability tasks and 13.09 min. for the retention tasks, so fatigue was not very relevant.
- Persistence effects. In order to avoid persistence effects, the experiment was carried out by subjects who had never done a similar experiment.
- Subject motivation. We motivated the students to participate in the experiment by offering them an extra point in the final mark of the course.

Another possible threat to the study's internal validity is *experimental bias*. Some students at the UPV may have perceived the RUP extension to be more useful than the other students and practitioners at the UCLM because they associate it with the RUP extension (which was developed at their university), and not because of their perceived performance in the experimental task. They may also have indicated a greater intention to use the RUP extension solely to please the experimenters. We believe we have reduced this threat by informing the students that their participation in the experiment would have no effect on their course mark.

Regardless of whether or not this control was effective, the UCLM data sets present 'independent' samples of participants, and the same pattern of results could be observed in these data sets. Furthermore, a different experimenter conducted these replications. As pointed out by Shull *et al.* [35], the independence of the replicators from the original experimenters lends additional confidence that the original results were not the result of experimental bias.

Another possible limitation is related to the external validity (i.e., the generalization of the findings) of this study - namely the use of a limited number of requirements modeling artifacts. Another limitation to the external validity might be the use of students as experimental subjects. However, the subjects who participated in the experiment were final-year Computer Science students and can be considered to be representative of novice users of requirements modeling approaches (see Section 5.2). Moreover, the experiment was also replicated with a group of practitioners.

Finally, a further limitation of our study is that the proposed evaluation method was validated in the evaluation of a single requirements modeling method (i.e., the RUP extension). We plan to conduct further experimentation to apply the proposed evaluation method in the evaluation of several requirements modeling methods. In fact, our evaluation method can be used both to evaluate the likelihood of acceptance of a particular method in practice and to compare two or more requirements modeling methods, to assess which is the most effective (i.e., is more likely to be accepted in practice). This can be done by applying the same hypotheses defined in our theoretical model to different *requirements modeling* methods and by then comparing the results obtained. This will imply the design of other experiments focusing on method comparison.

9. CONCLUSIONS AND FUTURE WORK

We have presented a method for evaluating the quality of requirements modeling methods based on user perceptions. This method consists of a theoretical model explaining the relevant dimensions of quality for requirements modeling methods, together with a practical instrument to measure these quality dimensions.

The method was tailored using as a basis the Method Evaluation Model. This is a model for evaluating Information Systems design methods that uses both aspects of method success: actual performance and likelihood of acceptance in practice. We basically defined performance-based variables (e.g., understandability efficiency and understandability effectiveness) as influencing factors for the perception-based variables (i.e., Perceived Ease of Use, Perceived Usefulness, and Intention to Use). The proposed theoretical model was empirically tested through a family of four experiments which was conducted to evaluate the likelihood of acceptance in practice of a RUP extension for modeling requirements.

The main results obtained from the analysis of the data gathered in the family of experiments reveal that:

- The majority of the participants found the RUP extension to be quite useful and easy to use in supporting the software requirements modeling;
- The majority of the participants were also very positive about the use of the RUP extension in the future;
- The actual performance of the participants in the experiments determined their positive perceptions;
- The positive perceptions determine the intention to use the RUP extension;

The whole experimentation work has allowed us to attain the primary goal of this paper which is to evaluate the usefulness of our proposed evaluation method to evaluate requirements modeling methods. Our proposed evaluation method was found to be a suitable tool for assessing the acceptance of requirements modeling methods based on user perceptions. Although we tested our evaluation model with a specific requirements method (i.e., the RUP extension) it can be used with any other requirements modeling methods to determine the likelihood of its acceptance in practice.

The family of experiments also allowed us to test the predictive and exploratory power of the evaluation model. The proposed evaluation model provides a theoretical foundation with which to explain the acceptance of requirements modeling methods in practice. The evaluation model not only considers the quality of the requirements modeling artifacts but also the perceptions of those people using the method. The empirical testing of the evaluation method revealed that, with the exception of the Perceived Ease of Use vs. Perceived Usefulness relationship (which was partially confirmed), all the other relationships in the theoretical model proposed for predicting the acceptance of requirements modeling methods were confirmed. Nevertheless, as future work we plan to investigate the influence of other performance-based and perception-based variables on predicting the acceptance of requirements modeling methods in practice.

From a research perspective, our study adds new insights into the problem of how to effectively evaluate requirements modeling methods and artifacts. We combined empirically-validated methods and theories (i.e., the Method Evaluation Model and the Cognitive Theory of Multimedia Learning) to explain how the quality of requirements modeling artifacts influence the quality of a requirements modeling method. We believe that our work contributes to the body of knowledge for the evaluation of requirements modeling methods based on theoretical foundations and empirical studies.

From a practical perspective, we are aware that this study provides only preliminary results on the usefulness of our evaluation model and the likelihood of acceptance in practice of the RUP extension. There is a need for more empirical studies with which to test the evaluation model in other settings. Nevertheless, this study has value as a pilot study to test the RUP extension before its deployment.

In future work, we plan to apply the proposed evaluation method to other requirements modeling methods (e.g., TROPOS). We particularly wish to apply the evaluation method in the comparison of the likelihood of acceptance of our RUP extension with other similar methods. This will be conducted by applying the evaluation method as a basis to define new controlled experiments. We also intend to study other theories that could suggest other factors that might influence the acceptance of requirements modeling methods in practice. We particularly aim to use the Unified Theory of Acceptance and Use of Technology (UTAUT) [37] to study the link between the user acceptance of the requirements modeling method and individual or organizational issues (e.g., job performance, social factors).

REFERENCES

[1] S. Abrahão, G. Poels, A Family of Experiments to Evaluate a Functional Size Measurement Procedure for Web Applications, Journal of Systems and Software 82(2) (2009) 253-269, Elsevier.

[2] S. Abrahão, E. Insfran, J.A. Carsí, M. Genero, M. Piattini, Evaluating the Ability of Novice Analysts to Understand Requirements Models, In: Proc. 9th International Conference on Quality Software, QSIC'09, Jeju, Korea, 2009 (short paper).

[3] H. Al-Subaie, T. Maibaum, Evaluating the Effectiveness of a Goal-Oriented Requirements Engineering Method. In: Proc. of the 4th International Workshops on Comparative Evaluation in Requirements Engineering, CERE'06, Minneapolis/St. Paul, Minnesota, USA, pp. 8-19.

[4] M. Ali Babar, D. Winkler, S. Biffl, S, Evaluating the Usefulness and Ease of Use of a Groupware Tool for the Software Architecture Evaluation Process, In: Proc. of the 1st International Symposium on Empirical Software Engineering and Measurement, ESEM'07, Madrid, Spain, pp. 430-439.

[5] V. Basili, F. Shull, F. Lanubile, Building Knowledge through Families of Experiments, IEEE Transactions on Software Engineering 25(4) (1999) 435–437.

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

[7] F. Bodart, A. Patel, M. Sim, R. Weber, Should Optimal Properties Be Used in Conceptual Modelling? A Theory and Three Empirical Tests. Information Systems Research, 12(4) (2001) 384-405.

[8] D.T. Campbell, D.W. Fiske, Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix, Psychological Bulletin 56 (1959) 81-105.

[9] B.H.C. Cheng, J. M. Atlee. Research Directions in Requirements Engineering. Workshop on the Future of Software Engineering at ISCE 2007, L. C. Briand and A. L. Wolf (Eds.), Minneapolis, USA, 2007, 285-303.

[10] M. Ciolkowski, F. Shull, S. Biffl, A family of experiments to investigate the influence of context on the effect of inspection techniques. In: Proc. of the 6th International Conference on Empirical Assessment in Software Engineering, EASE'02, Keele, UK, 2002, pp. 48–60.

[11] J.A. Cruz-Lemus, A. Maes, M. Genero, G. Poels, M. Piattini. The Impact of Structural Complexity on the Understandability of UML Statechart Diagrams, Information Sciences 180(14) (2010) 2209-2220.

[12] A.M. Davis, S. Overmeyer, K. Jordan, J. Caruso, F. Dandashi, A. Dirda, G. Kincaid, G. Ledeboer, P. Reynolds, P.Ta Sitaram, M. Theefanos, Identifying and measuring quality in a software requirements specification. In: Proc. of the 1st Int. Software Metrics Symposium, Metrics'93, 1993, pp. 141-152.

[13] F.D. Davis, Perceived Usefulness, Perceived Ease of Use and User Acceptance of Information Technology, MIS Quarterly 13(3) (1989) 319-340.

[14] M. Fishbein, I. Ajzen, Beliefs, Attitude, Intention and Behavior. An Introduction to Theory and Research. Reading, MA, 1975.

[15] M. Geisser, T. Hildenbrand, F. Rothlauf, C. Atkinson, An Evaluation Method for Requirements Engineering Approaches in Distributed Software Development Projects. In Proc of the Int. Conference on Software Engineering Advances, ICSEA'07, France, 2007, 39 - 49.

[16] A. Gemino, Y. Wand, Complexity and clarity in conceptual modeling: Comparison of mandatory and optional properties, Data Knowledge Engineering 55(3) (2005) 301-326.

[17] P. Haumer, IBM Rational Method Composer, http://www-128.ibm.com/ developerworks/rational/ library/jan06/haumer/ (April 2006).

[18] M. Höst, B. Regnell, C. Wholin, Using students as subjects—a comparative study of students and professionals in lead-time impact assessment, In: Proc. of the 4th Conference on Empirical Assessment and Evaluation in Software Engineering, EASE'00, Keele, UK, 2000, pp. 201–214.

[19] E. Insfran, O. Pastor, R. Wieringa, Requirements Engineering-Based Conceptual Modelling. Requirements Engineering 7 (2), (2002) 61–72.

[20] E. Insfran, (2003). A Requirements Engineering Approach for Object-Oriented Conceptual Modeling, PhD Thesis, DSIC, Valencia University of Technology, Spain.

[21] ISO, ISO/IEC 9126-1, Software Engineering – Product quality – Part 1: Quality model, 2001.

[22] I. Jacobson, G. Booch, J. Rumbaugh, The Unified Software Development Process, Addison-Wesley, 1999.

[23] B.A. Kitchenham, S. Pfleeger, D.C. Hoaglin, K. El Emam, J. Rosenberg. Preliminary Guidelines for Empirical Research in Software Engineering, IEEE TSE 28(8), (2002) 721–734.

[24] P. Kruchten, The Rational Unified Process: An Introduction, Addison-Wesley, UK, 1998.

[25] K. Mathieson, Predicting User Intention: Comparing the Technology Acceptance Model with the Theory of Planned Behaviour, Information Systems Research, 2(3) (1991) 173-191.

[26] R.E. Mayer, Multimedia Learning, Cambridge University Press, 2001.

[27] R.E. Mayer, Models for Understanding. Review of Educational Research, 59(1) (1989) 43-64.

[28] D.L. Moody, Dealing with complexity: a practical method for representing large entity relationship models, PhD Thesis, Dept. Information Systems, University of Melbourne, Australia, 2001.

[29] U. Nikula, J. Sajanie, Evaluation Framework for Requirements Engineering Method Adoption: The BaRE Method Case, In: Proc. of the 3rd International Workshop on Comparative Evaluation in Requirements Engineering, CERE'05, Paris, France, 2005, pp. 45-55.

[30] J. Nunally, Psychometric Theory, 2nd ed. ed. New York, NY, McGraw-Hill, 1978.

[31] K. Pohl, The three dimensions of requirements engineering: a framework and its applications, Information systems 19(3) (1994) 243-258.

[32] N. Rescher, The Primacy of Practice. Oxford, Basil Blackwel, 1973.

[33] L. Reynoso, M. Genero, M. E. Manso, M. Piattini, M. Assessing the Influence of Import-Coupling on OCL Expression Maintainability: A Cognitive Theory-Based Perspective. Information Sciences (2010), DOI 10.1016/j.ins.2010.06.028.

[34] C.K. Riemenschneider, B.C. Hardgrave, F.D. Davis, Explaining Software Developer Acceptance of Methodologies: A Comparison of Five Theoretical Models, IEEE Transactions on Software Engineering, 28(12) (2002) 1135-1145.

[35] F.J. Shull, J.C. Carver, S. Vegas, N. Juristo, The role of replications in Empirical Software Engineering, Empirical Software Engineering 13 (2008) 211-218.

[36] M. Turner, B. Kitchenham, D. Budgen, P. Brereton, Lessons Learned Undertaking a Large-scale Systematic Literature Review, In: Proc. of the 12th International Conference on Evaluation and Assessment in Software Engineering, EASE’08, Bari, Italy.

[37] V. Venkatesh, M.G. Morris, G.B. Davis, and F.D. Davis, User acceptance of information technology: Toward a unified view, MIS Quarterly 27 (2003) 425-478.

[38] Y. Wand, R. Weber, Information systems and conceptual modeling: a research agenda, Information Systems Research 13(4) (2002) 363–376.

[39] K. Wasson, A Case Study in Systematic Improvement of Language for Requirements, In: Proc. of the 14th IEEE International Requirements Engineering Conference, RE'06, Minneapolis, USA, 2006, pp. 9-18.

[40] D. Williams, M. Kennedy, A Framework for Improving the Requirements Engineering Process Effectiveness, Int. Council on Systems Engineering Conference, Brighton, UK, 1999.

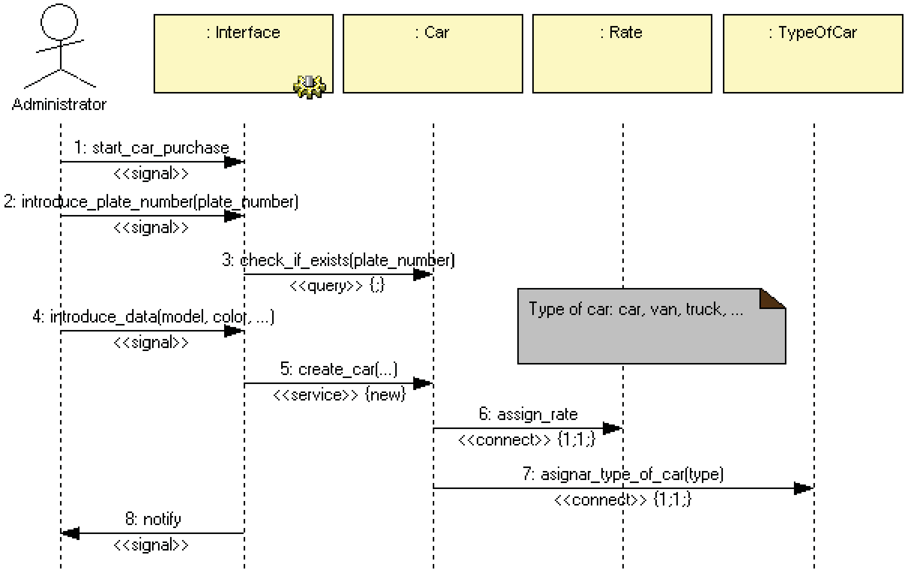
[41] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell, A. Wesslén, Experimentation in Software Engineering: An Introduction, Kluwer Academic Publishers, 2000.

Appendix A. Experimental Materials

As an example, the materials related to the rent-a-car domain are presented as follows.

A.1. An Example of Understanding Task (Test 1)

The following sequence diagram represents the creation of a new *Car* for a car rental company. All the company's cars have an assigned *Rate.* A type of car (car, van, truck, etc.) must also be assigned.



Answer the following Yes/No questions:

WRITE DOWN STARTING TIME (HH: MM: SS) ____________

1. Is it possible to create several *Cars* in this scenario? ________
2. Is the *Administrator* the person who can create a new rental *Car*? ________
3. Are there three classes in this Sequence Diagram? ________
4. If an appropriate *Rate* for a *Car* does not exist, can a new type of *Rate* be created and then assigned to the *Car*? ________
5. Is it possible to create a *Car* without a *Rate* (and later, before renting the car, indicate the corresponding rate)?________
6. Is it possible to have a *Car* which does not have a *Type of Car* assigned to it? ________

WRITE DOWN FINISHING TIME (HH: MM: SS) ____________

A.2. An Example of Retention Task (Test 2)

Fill the gaps in order to describe the functionality of the previous diagram.

WRITE DOWN STARTING TIME (HH: MM: SS) ____________

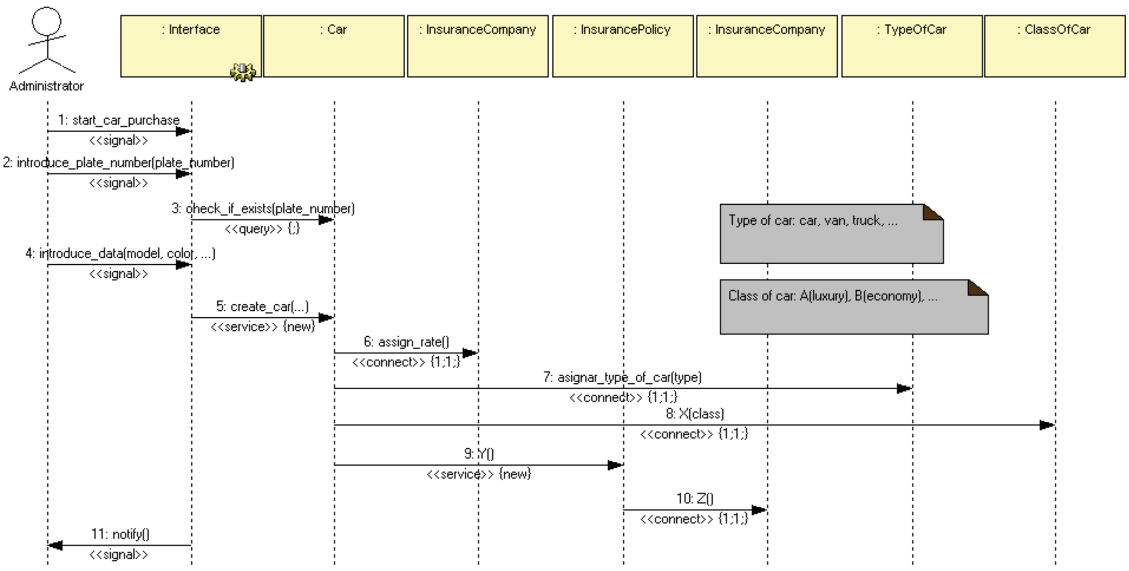
The following diagram represents the creation of a new car for the car rental company. The process begins when the _______ indicates to the interface of the system that he/she wants to create a new car. The Administrator introduces the _________ and the system _______ into the car rental fleet if the car already _______.

The Administrator then _______ the remaining car _____ (model, *color*, etc.) and, with all this information, uses the system to _______ a new car. One and only one rate is ______ to this new car, which is the rental price per day, and in addition, a type of car is assigned (car, van, truck, etc.), depending on the intended use. Finally, the system interface ______ to the _________ the creation of the new car.

WRITE DOWN FINISHING TIME (HH: MM: SS) ____________

A.3. An Example of Transfer Task (Test 3)

Consider the following functionality extension for the creation of a Car in the car rental system. When creating a new car, it is also necessary to indicate the class to which the car belongs (A = Luxury, B = Economy) and to buy an insurance policy from an insurance company.



Assign an appropriate name to messages 8, 9, and 10 (X, Y, and Z respectively).

WRITE DOWN STARTING TIME (HH: MM: SS) ____________

X = __________________________

Y = __________________________

Z = __________________________

WRITE DOWN FINISHING TIME (HH: MM: SS) ____________

Appendix B. An Example of inter-correlation analysis

Table B.1. Inter-item correlation analysis for the first experiment

		Perceived Ease of Use				Perceived Usefulness							Intention to Use				Overall		
		Q1	Q3	Q4	Q6	Q2	Q5	Q7	Q9	Q10	Q11	Q12	Q8	Q13	Q14	Q15	CV	DV	VALID?
PEOU	Q1	1.00	0.27	0.319	0.44	0.27	0.17	0.12	0.12	0.13	0.26	0.18	0.28	0.1	0.29	0.4	0.35	0.21	YES
	Q3	0.27	1	0.33	0.43	0.02	0.18	0.36	0.04	-0.05	-0.14	-0.06	0.33	0.15	0.36	0.26	0.34	0.18	YES
	Q4	0.32	0.33	1	0.57	-0.06	-0.08	-0.03	0.07	-0.17	0.06	0.02	0.2	-0.02	0.35	0.29	0.41	0.12	YES
	Q6	0.44	0.43	0.57	1	0.22	0.13	0.28	0.14	0.08	0.21	0.17	0.15	0.08	0.26	0.26	0.48	0.15	YES
PU	Q2	0.27	0.02	-0.06	0.22	1	0.27	-0.02	0.11	0.36	0.33	0.21	0.19	0.25	0.29	0.24	0.21	0.19	YES
	Q5	0.17	0.18	-0.08	0.13	0.27	1	0.21	0.53	0.33	0.27	0.31	0.44	0.01	0.3	0.43	0.32	0.22	YES
	Q7	0.12	0.36	-0.03	0.28	-0.02	0.21	1	0.17	0.17	0.18	0.45	0.32	0.13	0.27	0.21	0.19	0.22	NO
	Q9	0.12	0.04	0.07	0.14	0.11	0.53	0.17	1	0.45	0.48	0.24	0.15	-0.03	0.3	0.3	0.33	0.14	YES
	Q10	0.13	-0.05	-0.17	0.08	0.36	0.33	0.17	0.45	1	0.42	0.28	0.18	0.34	0.18	0.29	0.33	0.18	YES
	Q11	0.26	-0.14	0.06	0.21	0.33	0.27	0.18	0.48	0.42	1	0.49	0.15	-0.03	0.16	0.48	0.36	0.19	YES
	Q12	0.18	-0.06	0.02	0.17	0.21	0.31	0.45	0.24	0.28	0.49	1	0.38	0.14	0.42	0.48	0.33	0.23	YES
ITU	Q8	0.28	0.33	0.2	0.15	0.19	0.44	0.32	0.15	0.18	0.15	0.38	1	0.12	0.6	0.58	0.43	0.25	YES
	Q13	0.1	0.15	-0.02	0.08	0.25	0.01	0.13	-0.03	0.34	-0.03	0.14	0.12	1	0.31	0.11	0.18	0.12	YES
	Q14	0.29	0.36	0.35	0.26	0.29	0.3	0.27	0.3	0.18	0.16	0.42	0.6	0.31	1	0.68	0.53	0.29	YES
	Q15	0.4	0.26	0.29	0.26	0.24	0.43	0.21	0.3	0.29	0.48	0.48	0.58	0.11	0.68	1	0.63	0.33	YES