J Braz Comput Soc (2010) 16: 3–19
DOI 10.1007/s13173-010-0003-5

INVITED PAPER

An empirical comparative evaluation of requirements engineering methods

Sergio España · Nelly Condori-Fernandez · Arturo González · Óscar Pastor

Received: 22 February 2010 / Published online: 6 May 2010

Abstract Requirements Engineering (RE) is a relatively young discipline, and still many advances have been achieved during the last decades. In particular, numerous RE approaches are proposed in the literature with the aim of understanding a certain problem (e.g. information systems development) and establishing a knowledge base that is shared between domain experts and developers (i.e. a requirements specification). However, there is a growing concern for empirical validations that assess RE proposals and statements. This paper is related to the assessment of the quality of functional requirements specifications, using the Method Evaluation Model (MEM) as a theoretical framework. The MEM distinguishes the actual efficacy and the perceived efficacy of a method. In order to assess the actual efficacy or RE methods, the conceptual model quality framework by Lindland et al. can be applied; in this paper, we focus on the completeness and granularity of requirements models and extend this framework by defining four new metrics (e.g. degree of functional encapsulations completeness with respect to a reference model, number of functional fragmentation errors). In order to assess the perceived efficacy, conventional questionnaires can be used. A laboratory experiment with master students has been carried out, in order to compare (using the proposed metrics) two RE methods; namely, Use Cases and Communication Analysis. With respect to actual efficacy, results indicate greater model quality (in terms of completeness and granularity) when Communication Analysis guidelines are followed. With respect to perceived efficacy, we found that Use Cases was perceived to be slightly easier to use than Communication Analysis. However, Communication Analysis was perceived to be more useful in terms of determining the proper business processes granularity. The paper discusses these results and highlights some key issues for future research in this area.

Keywords Experiment · Requirements specification · Use Cases · Communication Analysis · Perceived usefulness · Perceived ease of use · Method Evaluation Model · Conceptual model quality framework

This paper revises and extends previous work that has been published in the 17th IEEE International Requirements Engineering Conference (RE'09) [1].

S. España (✉) · N. Condori-Fernandez · Ó. Pastor
Centro de Investigación en Métodos de Producción de Software, Universidad Politécnica de Valencia, Valencia, Spain
e-mail: sergio.espana@pros.upv.es

N. Condori-Fernandez
e-mail: nelly@pros.upv.es

Ó. Pastor
e-mail: opastor@pros.upv.es

A. González
Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, Valencia, Spain
e-mail: agdelrio@dsic.upv.es

1 Introduction

Requirements Engineering (RE), in spite of being a relatively young discipline, has achieved an ever growing body of knowledge. Numerous RE methods have been proposed over the last decades. Also, it is widely acknowledged that Requirements Engineering (RE) has a big impact in software quality [2]. However, most authors act as designers and propose new RE methods, while few authors act as real researchers validating that their (or other authors') proposals actually improve RE practice; although there is a growing concern for validating RE proposals, empirical evaluations are a strong need in the area [3, 4]. A grand challenge for

software research is to develop an understanding of which software methods work better and why [5], and it is likely that no standard method will suit all situations [6]. There exist works that address comparisons of RE approaches based on industrial experience [7]. However, seldom do we see work intended to compare features of RE approaches within an experimental context. This paper contributes to bridging the gap between the RE and the ESE (Empirical Software Engineering) communities.

In order to carry out method evaluations and comparisons, many evaluation techniques are available [8]. Some evaluation techniques are theoretical (e.g. ontological analysis) and others are empirical (e.g. laboratory experiments, action research); each type having its strengths and weaknesses. For instance, laboratory experiments provide high level of control and they are powerful in determining causality (internal validity is their strength), while the artificial environment in which they are carried out compromises the generalisation of the results (external validity is their weakness) [8, 9].

This paper presents a laboratory experiment that evaluates and compares two RE methods; namely Use Cases [10] and Communication Analysis [11]. Strictly speaking, the products of the following requirements specification (RS) techniques[1] are compared: namely, Use Case Diagrams and Communicative Event Diagrams (and their corresponding textual descriptions). These RS techniques allow the graphical and textual specification of what the literature usually refers to as functional requirements.

These two RE methods were chosen for the following reasons:

- Both methods have in common that they are applicable to the domain of information systems (ISs) development.
- The authors have academic and industrial experience of both methods.
- Use Cases is a widely known RE method that, although it is used in many industrial projects [13], it has also generated strong debates on their usefulness [14], numerous efforts to propose useful methodological guidelines have been made [15–18], and previous empirical works have been undertaken [19, 20].
- Communication Analysis is a novel RE method compared to Use Cases, but several Spanish companies already apply it successfully [11]. Also, it is soundly founded in systems theory and communication theory [21, 22]. This RE method places much emphasis in the guidelines that allow encapsulating functional requirements [23]. We expect Communication Analysis to perform better than Use Cases in practice.

In order to compare both RE approaches, we adopt (and adapt) the Method Evaluation Model (MEM), which defines a theoretical model and associated measurement instruments for evaluating information system design methods [24]. The MEM incorporates two aspects of method success; namely, *actual efficacy* (whether the method improves performance of the task) and *adoption in practice* (whether the method is used in practice). Additionally, in order to better understand practitioners' reaction to the method, the MEM also includes variables related to *perceived efficacy*. In this paper, we focus on actual efficacy and perceived efficacy.[2] With regard to actual efficacy, the conceptual model quality framework by Lindland et al. is applied. This framework distinguishes syntactic, semantic, and pragmatic quality. In this paper, we focus on semantic quality; the framework is extended with four new metrics that are aimed at the assessment of model completeness and model granularity. With regard to perceived efficacy, a post-task survey is used.

Many authors have theorised and empirically validated conceptual model completeness (the degree to which a model specifies all the relevant statements of a domain) [25, 26]. However, the approach often consists of a reviewer rating completeness on a Likert scale (or similar), and the procedure for assigning the rating depends on a subjective judgement [25, 27, 28]. We propose an approach based on the comparison of the reviewed model with a reference model that is built and agreed by an expert modelling committee. The metrics and the measuring procedure are precisely defined; for instance, the metric for completeness is not a Likert scale but an objective ratio that, in short, depends on the size complexity of the domain and the amount of information specified in the model.

Granularity, on its turn, is a much less investigated quality aspect, although it is an issue that has provoked debate in academic and industrial communities, particularly with regard to functional requirements such as use cases [29–31]. Kulak and Guiney define use case granularity as the relative scope of individual use cases compared to the application's scope [31]. In a more general sense, modularity is the design principle of having a complex system composed from smaller subsystems that can be managed independently yet function together as a whole [32] and granularity measures the size of encapsulations or modules (it is a systemic notion). When modelling, the analyst relies on methodological guidelines in order to encapsulate concepts in modelling primitives. We refer to the criteria that allow determining granularity as unity criteria [23]. It has been proved that

[1] We refer as method to a systematic way of working by which one can obtain a desired result. We refer as technique to a recipe for obtaining a certain result. It can be considered that methods contain techniques to perform particular tasks, and that techniques prescribe a way of working in detail (these definitions are borrowed from [12]).

[2] Measuring adoption in practice implies follow-up studies on the experimental subjects and this was not possible in our setting.

modularity improves business process model understanding [33]. This paper proposes an evaluation of the quality of modularity that goes beyond previous proposals since sound unity criteria for business process modelling are taken as reference and precise metrics for granularity errors are defined.

In any case, for an RE method to be successful, it is not enough that the method is theoretically sounder, not even that it performs better in practice, but it also needs to be perceived by practitioners as more useful and easier to use [24].

The contributions of the paper are the following:

- The paper discusses the concepts of completeness and granularity from a semantic point of view and grounds them in sound theory.
- The concepts of completeness and granularity are operationalised by proposing metrics for their quantification. Some metrics extend previous approaches; other metrics adopt a novel approach that, according to the results, is promising.
- The paper presents an experiment that compares two functional RS techniques (Use Case Diagram and Communicative Event Diagram), integrating the above-mentioned concepts and metrics into the method evaluation model. Although we focus on two particular methods, the proposed strategy is general enough to fit the evaluation of other methods.

Evidence shows that Communication Analysis leads to models that are more complete and have less modularity errors. Also, though Use Cases is perceived to be easier to use, Communication Analysis is perceived to be more useful. These outcomes are discussed and related to the qualities of the techniques.

The paper is structured as follows. Section 2 presents the theoretical framework for method comparison, which consists of an integration of the MEM and the conceptual model quality framework by Lindland et al., extended with new metrics. Section 3 reviews previous evaluations of RS techniques. Section 4 overviews the RS techniques evaluated in the laboratory experiment; namely, Use Case Diagram and Communicative Events Diagram. Section 5 presents the experimental planning. Section 6 analyses the results of the experiment. Section 7 discusses the validity of the results. Section 8 presents conclusions and future work.

2 Theoretical framework for the assessment of RS methods

2.1 The Method Evaluation Model (MEM)

Following Moody's approach to validating information system design methods [24], in deciding how to validate requirements specification techniques, one needs determining whether a method is successful or not. Moody argues that there are (at least) two dimensions of "success" that need to be considered in evaluating IS design methods:

- *Actual efficacy*, i.e. does the method improve performance?
- *Adoption in practice*, i.e. is the method used in practice?

Moody developed the Method Evaluation Model (MEM), a theoretical model that combines aspects of the Rescher's pragmatic method [34] and the Davis's Technology Acceptance Model [35].

As shown in Fig. 1, MEM includes six primary constructs: actual efficiency, actual effectiveness, perceived ease of use, perceived usefulness, intention to use, and actual usage.

Moody's model is based on the theory that *actual efficiency* and *effectiveness* determine *intentions to use* a

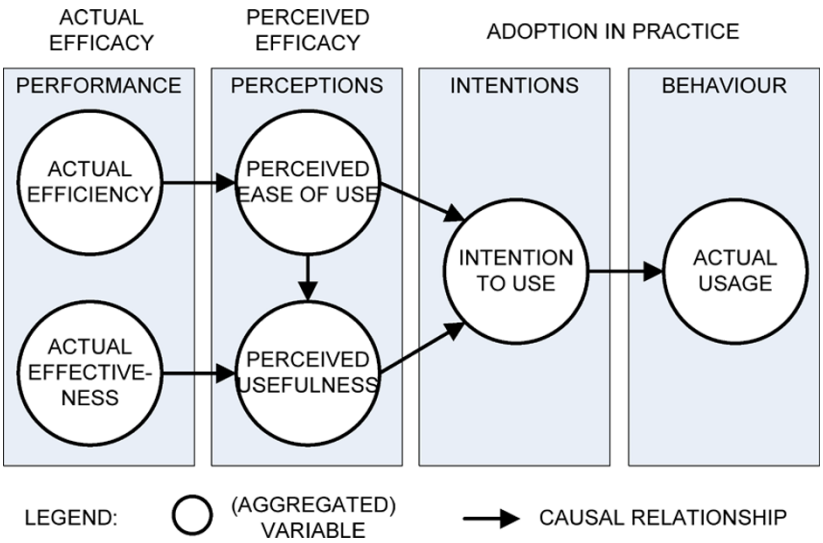

Fig. 1 Method evaluation model [24]

method only 'second-hand', via the *perceived ease of use* and the *perceived usefulness*. This is due to the fact that, in human behaviour, subjective reality is more important than objective reality. This paper is concerned with evaluating the RE methods with regards to the MEM, except for the actual usage, which is out of the scope.

2.2 Conceptual model quality frameworks

Lindland et al. [26] present a conceptual model quality framework that is founded on semiotics and linguistics. Three types of model quality are assumed:

- *Syntactic quality*. The degree to which the model adheres to the modelling language rules. Syntactic errors and deviations from the rules decrease syntactic quality.
- *Semantic quality*. The degree to which the model represents the domain. The more similar the model and the domain, the better the semantic quality.
- *Pragmatic quality*. The degree to which the model is correctly interpreted by its audience. The less misunderstanding, the better the pragmatic quality.

For each type of quality, absolute goals are defined and the means to achieve the goals are described. See Fig. 2 for an overview of the framework.

Other conceptual modelling quality frameworks have been proposed in the literature. Yadav et al. [25] propose a framework that provides criteria to compare information systems RE methods in terms of the modelling process and the model itself; criteria are classified in four categories: syntactic, semantic, usability, and communicating ability. Davis et al. [36] explore the concept of RS quality and consider completeness, but acknowledge that the proposed metrics are "difficult to measure". Pohl [37] develops a framework for RE with three dimensions (specification, representation, and agreement) and defines three goals for a RS (formally represented, complete, and agreed). In [38], the framework by Lindland et al. is extended by considering additional levels of the semiotic ladder [39]. In [40], the framework by Lindland et al. is extended with notions on social construction theory from Pohl's framework. Moody et al. [41] present a quality framework for data models oriented towards practice. Schuette et al. [42] present the Guidelines of Modelling, which is a framework of principles that improve the quality of information models by reducing subjectivism in the information modelling process.

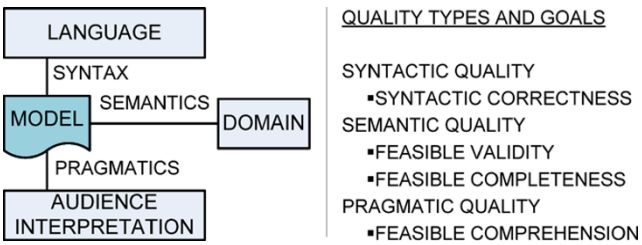

Fig. 2 Conceptual model quality framework [26]

Some reviews of the state of the art and framework comparisons can also be found. In [43], the frameworks by Krogstie, Moody et al., and Schuette et al. are compared by means of their metamodels. In [44], several frameworks are compared and the frameworks by Lindland et al. and Moody et al are integrated. In [45], an exhaustive review of related works is presented, and recommendations for research in conceptual model quality are given.

We have chosen the framework by Lindland et al. as a point of departure for our research for it has been acknowledged as a reference framework by many authors [38, 45] and it has been empirically tested and used for evaluation purposes [27, 28].

2.3 Adopting and extending MEM for evaluating requirements specification techniques

In the experiment described in Sects. 6 to 7, the MEM has been adopted as a framework for method comparison (see Fig. 3). As recommended in [24], actual efficacy variables have been further refined and adapted taking into account the particularities of the methods being compared and the task being evaluated, and a questionnaire has been used to measure perceived efficacy.

2.3.1 Actual efficacy

Lindland et al. [26] define that a model (M) has achieved semantic completeness if it contains all the statements about the domain (D) that are correct and relevant. That is, $D \backslash M = \varnothing$. However, except for extremely simple and highly inter-subjectively agreed domains, total completeness cannot be achieved (due to resource restrictions). Hence, they relax the completeness goal by applying the notion of feasibility. Feasibility introduces a trade-off between the benefits and drawbacks for achieving a given model quality.

A model has achieved *feasible completeness* when there is no relevant statement about the domain, not yet included in the model, such that the additional benefit to the conceptual model from including the relevant statement exceeds the drawbacks of including it. That is, $D \backslash M = S \neq \varnothing$, where S is the set of correct and relevant statements not yet in the model.

However, from a constructivist stance (such as in [38]), to determine whether there is a relevant statement of the domain not yet included in the model, one must first conceptualise the domain. This conceptual model of reference (Mr) needs not be written but at least it must exist in the reviewer's mind.

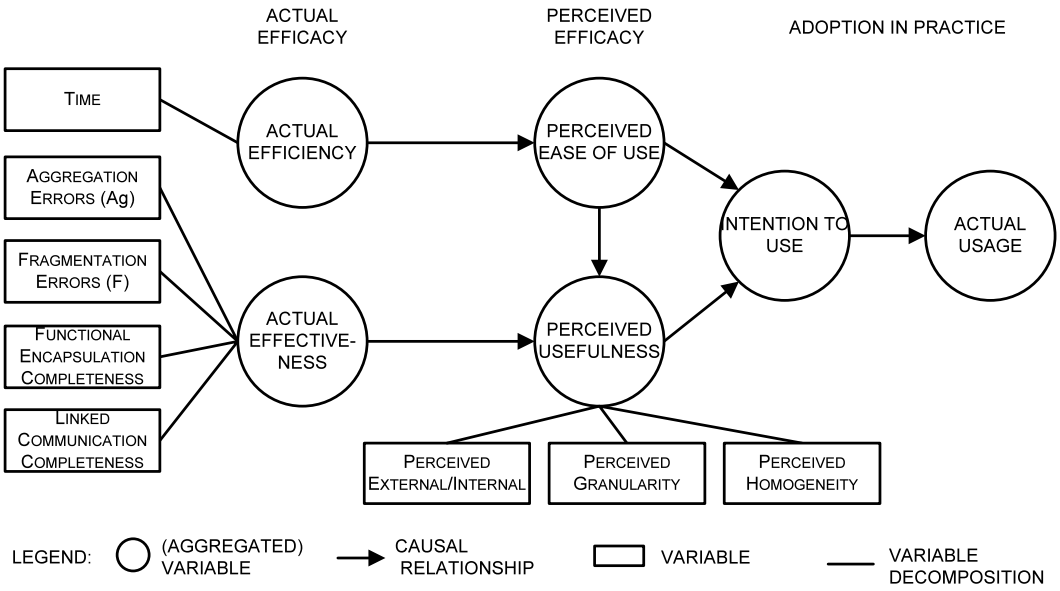

Fig. 3 Refined MEM used in the experiment

Also, Lindland et al. recommend adapting and refining the framework depending on the method being evaluated, as well as operationalising the framework by proposing metrics. For instance, they propose to decompose feasible completeness into feasible functional completeness, feasible non-functional completeness, etc. A germinal definition states that a model has achieved *feasible functional completeness* whenever all relevant functional requirements that are worth being specified have been included in the model (adapted from [26]). We further develop this idea by refining it, by explicitly considering the existence of a reference model, and by proposing metrics.

When Lindland et al. propose refining the framework they consider the existence of different types of statements about the domain, what leads to building models with multiple views. Each view can itself be considered a model. For instance, if the domain is considered to contain two types of statements (those about functionality—*FD*—and those about qualities and restrictions—*NFD*—), then the requirements model M will have two views: a functional requirements model FM and a non-functional requirements model NFM; that is, $D = FD \cup NFD$ and $M = FM \cup NFM$. Feasible functional completeness can now be formally defined as $FD \backslash FM = FS \neq \varnothing$, where FS is the set of functional requirements not yet in the model (but which are not worth the effort to be included).

Feasible functional completeness can be further refined if a specific type of domain is considered (e.g. information systems). Research about the essence of information systems has lead to the definition of sound conceptual frameworks. An information system (IS) is a socio-technical system that supports organisational communications [46]. The FRISCO report [39] laid the foundations of the area, upon which other researchers have built theories [21]. Based on this conception of ISs, two major abstract modelling primitives for functional RS are identified; namely functional encapsulations and linked communications.

- Wieringa defines function as a service provided by the IS to its environment; it is also considered to be an encapsulation of a useful behaviour of the system [12]. We therefore refer as *functional encapsulations* to IS functions, in order to highlight the importance of determining the boundaries of the encapsulation.
- We refer as *linked communication* to the message conveyance that is triggered by the occurrence of an event (the use or activation of a function) and by which the IS informs an actor of this occurrence.

This way, it is considered that a functional requirements model is composed of (at least) a set of functions (F) and a set of linked communications (LC). Then $FM = F \cup LC$. Figure 4 shows an example of these abstract primitives and Sect. 4 establishes their correspondence with the evaluated RS techniques.

Furthermore, a constructivist approach takes us to define completeness with respect to a reference model. In industrial settings, a reference model may be impractical; customer reviews are essential [36]. In experimental settings, we propose that an expert modelling committee analyse a given domain and agree a model that strictly follows best practices in modelling.[3] Some best practices depend on the modelling technique (e.g. correct use of use case *includes*

[3]This agreement may involve debate and several iterations.

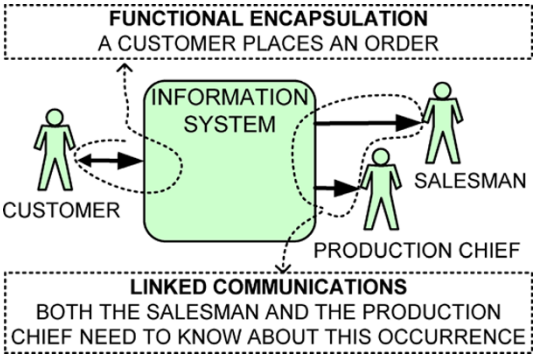

Fig. 4 Abstraction of the primitives of ISs functional requirements models

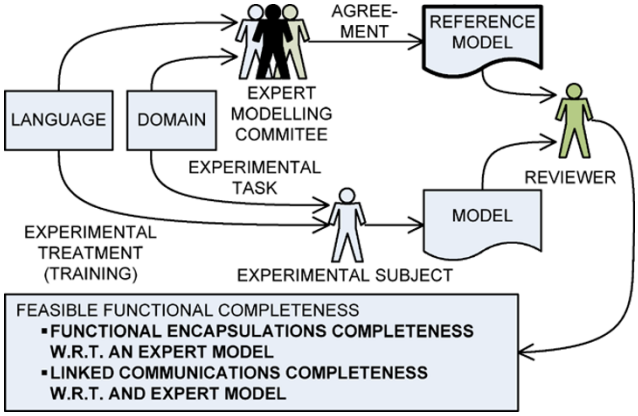

Fig. 5 Model completeness evaluation with respect to a reference model

and *extends* relations [10]) whereas others are independent (e.g. RS should provide an external view of the system [47]). The reference model is defined as $Mr = FMr \cup NFMr$, and the reference functional requirements model is defined as $FMr = Fr \cup LCr$. It is assumed that FMr contains all relevant functional requirements and linked communications.

Measurable quality goals are defined in terms of the reference model (see Fig. 5):

- *Functional encapsulations completeness with respect to (w.r.t.) a reference model*. All functional requirements specified in the reference model have been specified in the model. That is, $Fr \backslash F = \varnothing$. Note that the reference model substitutes the domain in the comparison.

This goal is related to the degree with which the reviewed model contains the functional encapsulations (i.e. use cases, communicative events) included in the reference model. A metric for this goal is the degree of functional encapsulations completeness w.r.t. a reference model, which is defined as $degFEC = |F|/|Fr|$.

- *Linked communications completeness w.r.t. a reference model*. All linked communications specified in the reference model have been specified in the model. That is, $LCr \backslash LC = \varnothing$.

This goal is related to the degree with which the reviewed model contains the linked communications (i.e. communications triggered by the occurrence of a use case or a communicative event) specified in the reference model. A metric for this goal is the degree of linked communications completeness w.r.t. a reference model, which is defined as $degLCC = |LC|/|LCr|$.

Assuming that it is feasible to build the reference model, both goals are feasible quality goals.

With regard to granularity, we claim that it is a quality of requirements models that has not been sufficiently investigated. We refer as *unity criteria* to the norms that guide the identification of complex concepts (e.g. use cases) and the encapsulation of their components (i.e. the flow of actions a use case is composed of); therefore, unity criteria determine the granularity of encapsulations [23]. Data models unity criteria are frequently based on the notion of identification; there is much consensus in this area. However, unity criteria for functional models (e.g. use case models) are far from being widely agreed. We argue that this lack of agreement leads to the proliferation of methodological guidelines that avoid the thorny issue of granularity or, in many cases, they define simplistic unity criteria (e.g. the *twenty use cases per system* rule of thumb [48], the *one person, one sitting* test [10]). Consequences are inconsistencies in modelling practice, heterogeneous model granularity, k and the need for gurus to whom consult [49].

In [23], we unfold the notion of unity criteria and we propose unity criteria for business process modelling. Industrial practice has shown us that the criteria reduce subjectivity in the encapsulation of business processes.

A model is considered to have an *appropriate granularity* with respect to a given unity criteria when the encapsulations of the model (typically under the form of modelling primitives) are conform to the unity criteria. Taking a given set of unity criteria it is possible to identify granularity errors in a model (see Fig. 6). If the reference model conforms to the unity criteria, it aids the review; but it is not strictly required.

For functional requirements models we propose identifying the following granularity errors:

- *Functional fragmentation error*. This error is the result of modelling two or more functional encapsulations for a part of the domain which, according to the given unity criteria, should have been modelled as only one encapsulation. For instance, two or more use cases of the reviewed model correspond to one use case of the reference model. It is measured as a variable named *errFra*.
- *Functional aggregation error*. This error is the result of modelling certain part of the domain as one functional en-

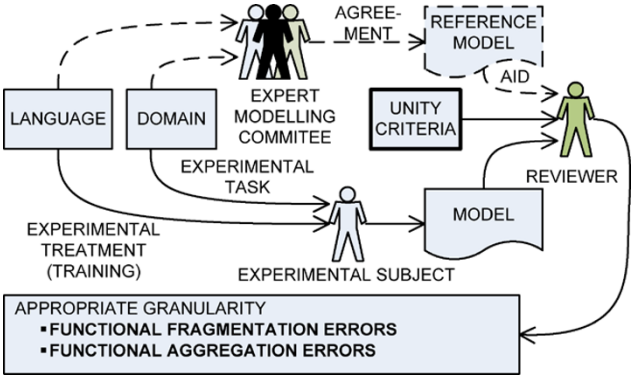

Fig. 6 Model granularity evaluation with respect to unity criteria

capsulation when, according to the unity criteria, the phenomena should be modelled using two or more encapsulations. For instance, two or more use cases of the reference model are modelled as only one use case in the reviewed model. It is measured as a variable named *errAgg*.

Therefore, a model can be said to achieve appropriate granularity whenever there are no granularity errors. That is, $errFra = 0$ and $errAgr = 0$.

2.3.2 *Perceived efficacy*

- *Perceived ease of use*: the degree to which a person believes that using a particular specification technique would be free of effort.
- *Perceived usefulness*: the degree to which a person believes that a particular specification technique will be effective in achieving its intended objectives. Three objectives were identified in our context to evaluate this perception:
  - Distinction between external and internal interaction;
  - Homogeneity of functional specification;
  - Adequate level of granularity of functional specification.

Next section, we present an experimental study, which was designed and reported following the recommendations provided by Juristo et al. [50] and Wohlin et al. [9].

3 Previous empirical evaluations of RS techniques

Some empirical works assess the quality of functional requirements models. Special interest has been placed in evaluating use case specification guidelines. In regard to completeness, two approaches can be distinguished: assessing the completeness of a (single) use case description, rating the completeness of a use case model based in the reviewer's judgement, and measuring the size of the use case model.

Some works focus on a single use case and analyse its detailed textual description. For instance, an experiment by Ben Achour et al. [51] lays emphasis on the specification of the use case flow of actions. Cox and Phalp [52] replicate the previous experiment and extend the marking scheme with subjective metrics. Instead, as explained in Sect. 2.3, we are interested in assessing the whole functional RS.

Other works focus on the whole use case model but rate completeness in terms of a value judgement. Yadav et al. [25] propose assessing completeness by having an expert review committee rate each model on a 1 to 7 scale, based on judgement. In experiments by Moody et al. [27, 28], quality ratings are also given on a 7 point Likert scale, from 1 (poor) to 7 (excellent). We advocate using metrics rather than ratings.

In [53], the authors use a 0 to 3 scale to rate several properties of the model (mainly related to correctness). In regard to completeness, the paper states that "the number of identified actors and use cases [...] indicate quality of the guidelines—the higher number, the better quality". We believe that this statement is only sensible if just valid actors and use cases are counted and if use cases have an appropriate granularity, but the paper does not give a more detailed explanation.

Fortuna et al. [54] describes an experiment that considers functional coverage and granularity homogeneity in order to validate unity criteria for Use Cases; however, the procedure is not sufficiently described to learn lessons from it.

4 Communication Analysis and Use Cases

4.1 Use Cases

Use Cases is an RE method proposed by Jacobson [55] and later revised by many authors [18, 56]. "A use case is a collection of possible sequences of interactions between the system under discussion and its external actors, related to a particular goal" [18]. Use Cases have been used in many industrial projects [13] but they have also generated strong debates on their usefulness [14]. Use Cases can be modelled graphically by means of the Use Case Diagram. Actors represent users that interact with the system under development. With regard to a use case an actor can play either a primary actor or a secondary actor role. Extension, inclusion, and generalisation relations can be established among use cases. Figure 7 shows some of the modelling primitives of the Use Case Diagram technique (see [56] for more detail); the figure also shows the relation between the Use Case Diagram technique and the abstract primitives. *Functional encapsulations* correspond to use cases. It is by means of a use case that the functionality of the system is encapsulated; in other words, use cases are modules that represent functions. For

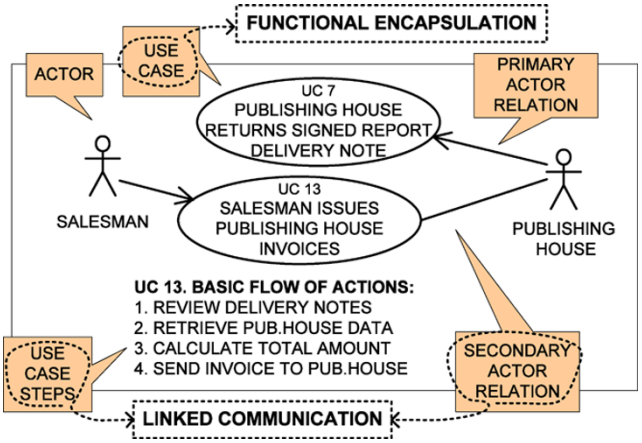

Fig. 7 Use case diagram fragment

instance, "UC 7 Publishing house returns signed report delivery note" encapsulates functionality related to the organisational work practice. *Linked communications* typically appear in the use case description (e.g. as a step in the flow of actions), but may as well appear as a relation between a use case and an actor. For instance, the fact that an invoice needs to be sent to a publishing house whenever the salesman issues monthly invoices is represented in two different ways: the use case "UC 13" has a secondary actor relation with "Publishing house" (which means that publishing houses are involved somehow in that use case) and step 4 in the "UC 13" flow of actions clarifies the type of communication.

Many works propose guidelines for use case modelling [16, 17]; guidelines by Cockburn [10] were chosen for the experiment because he explicitly proposes unity criteria that are based on user goals. Following Cockburn's methodological guidelines, the requirements engineer needs to take the following steps [10]:

1. Enumerate and define the actors involved.
2. Identify actor goals (see related guidelines below).
3. Make a list of use cases and a use case diagram, based on the list of goals defined in step 2.
4. Identify chances of reuse and update the diagram, making use of inclusion and extension relations.
5. For each use case, create a use case description template; specify the header fields (e.g. use case name, primary actor, preconditions) and give a brief textual description of the use case.
6. Define the basic flow of events and the alternative flows of actions.

Cockburn distinguishes several levels of goals to which use cases are related. The level of greatest interest is the user goal, i.e. the goal of the primary actor trying to get work done. In order to assess whether a goal is a user goal, Cockburn offers as guidelines two questions and a heuristic [10] (we consider them to be unity criteria for encapsulation):

- "Can you go away happy after having done this?"
- "Does your job performance depend on how many of these you do today?"
- A user goal passes the "one person, one sitting (2–20 minutes)" test.

4.2 Communication Analysis

Communication Analysis is an RE method that proposes undertaking IS analysis from a communicational perspective [11]. The method stems from ISs foundations research [21] and it evolves by means of the collaboration between industry and academia. Communication analysis is currently being applied to big projects in industrial environments; e.g. the integration of Anecoop S. Coop (a Spanish major distributor of fruit and vegetables) with its associated cooperatives (>100). Experience has shown us that the method can be grasped by practitioners and successfully put in practice.

Communication Analysis offers a requirements structure and several modelling techniques for requirements specification [11]. Among these techniques, we choose the Communicative Event Diagram because it is comparable to the Use Case Diagram. The Communicative Event Diagram is intended to describe business processes from a communicational perspective. A communicative event is a set of actions related to information (acquisition, storage, processing, retrieval, and/or distribution), which are carried out in a complete and uninterrupted way, on the occasion of an external stimulus. For each event, the actors involved are identified, as well as the corresponding communicative interactions and the precedence relations among events. Figure 8 shows the some of the modelling primitives of the Communicative Event Diagram technique (see [11] for more detail); the figure also shows the relation between the notation and the abstract primitives. *Functional encapsulations* correspond to communicative events. For instance, communicative event "PHO 7 Publishing house returns signed report delivery note" encapsulates part of the functionality of the system. *Linked communications* correspond to outgoing communicative interactions [11]. Unlike the Use Case Diagram technique, the Communicative Event Diagram has a specific modelling primitive devoted to linked communications. For instance, the outgoing arrow from "PHO 13" to the actor "Publishing house" indicates that, whenever the salesman issues an invoice, the invoice has to be sent to the publishing house (it represents a message conveyance).

Following Communication Analysis guidelines, the requirements engineer needs to take the following steps:

1. Define the purpose of the organisation, identifying the main business objects and the main functions the users want to apply to each business object.

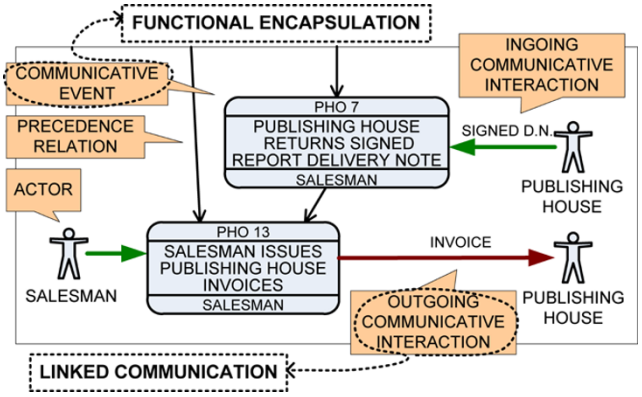

Fig. 8 Communicative event diagram fragment

2. For each business object, identify the initial communicative event that creates the object.
3. Assign an identifier and give the communicative event an appropriate name.
4. Define the encapsulation of the communicative event (see unity criteria below).
5. Identify other communicative events that affect the business object under analysis. For each new communicative event repeat steps 3 and 4.
6. Identify outgoing communicative interactions (i.e. output messages) that users need to carry out their work.

Communication Analysis offers the following unity criteria, which serve as guidelines for identifying (encapsulating) and modelling communicative events [23]:

- *Trigger unity*. The event occurs as a response to an external interaction and, therefore, some actor triggers it.
- *Communication unity*. Each and every event involves providing new meaningful information. This input message needs to be specified.
- *Reaction unity*. The event is a composition of synchronous activities. Events are asynchronous among each other.

The modelling notation of both techniques is quite similar. However, their modelling primitives have distinct underlying concepts and, therefore, the unity criteria are different.

5 Experimental planning

The goal of our experiment, according to the goal/question/metric template [57], is to analyze functional requirements specifications (RS) with the purpose of carrying out a comparative evaluation of RS methods with respect to their actual effectiveness and perceived efficacy from the viewpoint of the researcher in the context of computer science master students.

With respect to the actual effectiveness of the RS methods, the experiment addresses the following research questions:

RQ1: Will the subjects applying Communication Analysis produce functional RS with higher degree of completeness than the subjects applying Use Cases?

RQ2: Will the subjects applying Communication Analysis produce functional RS with a more appropriate level of granularity than the subjects applying Use Cases?

With respect to the perceived efficacy of the RS methods, the experiment addresses the following research questions:

RQ3: Will Communication Analysis be perceived as easier to use than Use Cases?

RQ4: Will Communication Analysis be perceived as more useful than Use Cases?

5.1 Experimental context

The selected subjects were 36 computer science master students enrolled in the 2007–2008 "Conceptual Modelling of Information Systems" course at Universidad Politécnica de Valencia, Spain. Participation was anonymous (aliases were used instead of names) and students were ensured that performance would not influence academic marks.

Previous to this course, students have taken several courses on programming, design and analysis (a bottom-up pedagogical approach is followed), so they have knowledge on data structures and algorithms, organisations and information systems, structured analysis and design, object-oriented analysis, database design and technologies, and requirements engineering principles (see the Faculty of Computer Science website for more details on the degree http://www.fiv.upv.es/default_i.htm).

With the purpose of identifying the background and experience using different specification techniques, such as Data Flow Diagram, Use Cases Diagram, Activity Diagram, Entity Relationship Diagram, Class Diagram, and Workflow Diagram; a demographic questionnaire was applied. 62.5% of the students answered to have a good knowledge about the syntaxes of the Entity Relationship Diagram (score superior to 3 points using a 5-likert scale). Experience using this technique was also good since 45% reported having to solve complex case studies. However, 83% of the students had little knowledge about the syntaxes of Use Case Diagram (score inferior to 3 points using a 5-likert scale). Similar results were obtained for other specification techniques. Therefore, prior to the experimental task, the subjects were trained to use adequately the respective guidelines of Use Cases and Communication Analysis.

As we want to compare two treatments (that is, both RE methods) against each other; a paired comparison design [9] was planned: each subject uses both treatments on the same

Table 1 Paired comparison design

Subjects	Treatment	
	(A)	(B)
	Communication Analysis	Use Cases
Group 1	1st round	2nd round
Group 2	2nd round	1st round

object; to minimise the effect of the order in which subjects apply the treatments, both orders need to be considered. As Table 1 shows, the experiment was carried out in two groups that were formed according to student's availability; each subject applied each of both methods (treatment) to specify the requirements of a photography agency IS (the experimental object). This was enacted in two rounds.

The experimental object is a four pages textual description of the needs for an IS that supports the work practice of a photography agency (it also includes some organisational forms). This enterprise acts as an intermediary between photographers that provide illustrated reports and publishing houses that request and buy those reports.

Using two different experimental objects (two IS descriptions) would have avoided the fact that subjects already know the problem domain when they face the second round. However, we chose to use only one experimental object due to timing constraints (each round lasted several weeks) and also to avoid tiredness effects in the subjects that would affect their perceptions of the methods.

It should be noted that, while similar experiments have used as experimental object a single use case [51, 52], our intention was to assess the guidelines of the respective methods for dealing with complete ISs, so we needed a bigger experimental object. The reference model (built by a committee of expert modellers) has 13 use cases; its corresponding software application (built by a software development company) has 537 function points.

5.2 Variables

We identified three types of variables [50]:

Response variables. In our study, functional completeness (functional encapsulation completeness, linked communication completeness), granularity (aggregation errors, fragmentation errors), perceived ease of use (PEOU), and perceived usefulness (PU) were identified as outcomes of the experiment. Section 2.2 defines the metrics: *degFEC*, *degLCC*, *errFra*, and *errAgg* for the first two response variables. The latter two variables were measured using a 5-point Likert scale format to gather users' perceptions.

Factors. The RE method was identified as a variable that could affect the response variables. Two treatments were considered: (1) Use Cases (mainly Use Case diagram and textual use case descriptions) and (2) Communication Analysis (mainly Communicative event Diagram). Another factor is the group to which the subject belongs, since each group applied the methods in a different order.

Parameters. Variables that we do not want to influence the experimental results have been fixed: application domain, complexity of the IS (problem statement) and previous RE experience.

5.3 Hypotheses

The hypotheses formulated from the research questions defined above are the following:

Hypothesis 1 Null hypothesis, H_{10}. Use Cases and Communication Analysis allow obtaining RS with same degree of functional encapsulations completeness.

Alternative hypothesis, H_{11}. Communication Analysis allows obtaining RS with greater degree of functional encapsulations completeness than Use Cases.

Hypothesis 2 Null hypothesis, H_{20}. Use Cases and Communication Analysis allow obtaining RS with same degree of linked communications completeness.

Alternative hypothesis, H_{21}. Communication Analysis allows obtaining RS with greater degree of linked communications completeness than Use Cases.

Hypothesis 3 Null hypothesis, H_{30}. Use Cases and Communication Analysis allow obtaining RS with same number of functional fragmentation errors.

Alternative hypothesis, H_{31}. Communication Analysis allows obtaining RS with less functional fragmentation errors than Use Cases.

Hypothesis 4 Null hypothesis, H_{40}. Use Cases and Communication Analysis allow obtaining RS with same number of functional aggregation errors.

Alternative hypothesis, H_{41}. Communication Analysis allows obtaining RS with less functional aggregation errors than Use Cases.

Hypothesis 5 Null hypothesis, H_{50}. Use Cases and Communication Analysis are equally perceived as easy to use.

Alternative hypothesis, H_{51}. Communication Analysis is perceived as easier to use than Use Cases.

Hypothesis 6 Null hypothesis, H_{60}. Use Cases and Communication Analysis are equally perceived as useful.

Alternative hypothesis, H_{61}. Communication Analysis is perceived as more useful than Use Cases.

5.4 Instrumentation

The instruments used in this experiment include the demographic questionnaire, the experimental object, training materials, and the post-task survey.

The experimental object is a problem statement that describes in natural language the structure and business processes of a photography agency.

The training materials are the following: a set of instructional slides on Use Cases guidelines based on the work of Cockburn [10], and Communication Analysis guidelines based on the work of Gonzalez et al. [11]. These guidelines are summarised in Sect. 4.

The survey instrument includes thirteen closed questions (5-point Likert scale) that were identified for measuring response variables. Perceived ease of use was measured using 6 items in the survey (Questions 1, 2, 4, 6, 10, and 12); perceived usefulness was measured using 7 items in the survey (Questions 3, 5, 7, 8, 9, 13, and 14).

The demographic questionnaire consisted on 26 questions (5-point Likert scale) aimed at assessing the subjects level of knowledge on several IS analysis and design methods; their actual experience with those methods; their appreciation on how important those methods are in industrial projects; whether, to the best of their knowledge, there exist enough methodological support and guidelines for applying those methods; the size complexity of software they have programmed before; and the degree of modularity of the software they develop.

5.5 Experimental procedure

The experimental procedure is depicted in Fig. 9. The experiment was initiated with the subjects training on two RE methods. The time used for training in each RE method was 8 hours distributed over 4 days.

During the first round, Group 1 was trained in Communication Analysis, and Group 2 was trained on Use Cases. Then the subjects received a natural language problem statement describing the structure and processes of a photography agency. They applied the RS method to specify the needed IS. The second round was analogous but now Group 1 was trained on Use Cases, and Group 2 was trained on Communication Analysis.

In order to capture the subjects' first impressions after using the respective techniques to specify the photography agency system, they completed a survey. The collected data was used to evaluate the perception-based variables.

6 Results analysis and interpretation

6.1 Actual effectiveness of the RS methods

An expert reviewer used the reference model and the unity criteria to aid him in the correction of the RS models. The measures obtained by the expert reviewer were analyzed. See Table 2 (a suffix is added to the variable name to indicate the technique: CA stands for Communication Analysis, UC stands for Use Cases).

The degree of functional encapsulations completeness w.r.t. a reference model (*degFEC*) is 81% for Communication Analysis and 75.64% for Use Cases. A bigger difference appears in the degree of linked communications completeness (*degLCC*), with up to 75% for Communication Analysis and only 50.46% for Use Cases. With regard to granularity errors, both error measures indicate that subjects applying Communication Analysis perform better than subjects applying Use Cases.

Table 2 Descriptive statistics

Measure	*N*	Mean	Std. deviation
degFEC_CA	34	0.8100	0.11710
degFEC_UC	36	0.7564	0.11557
degLCC_CA	34	0.7500	0.22937
degLCC_UC	36	0.5046	0.15164
errFra_CA	34	1.8824	0.97746
errFra_UC	36	2.1944	1.75368
errAgg_CA	34	0.5588	0.78591
errAgg_UC	36	1.2222	0.68080

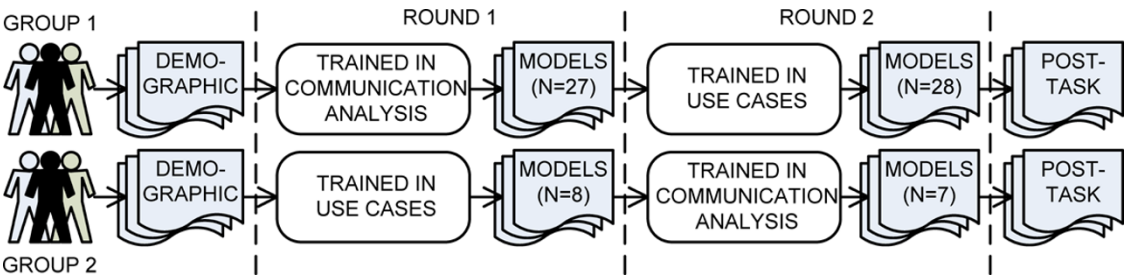

Fig. 9 Experimental procedure

Table 3 Paired samples test for *degFEC* and *degLCC* measures

	95% confidence interval of the difference		*t*	Sig. (2-tailed)
	Lower	Upper		
degFEC_CA − *degFEC_UC*	−0.00178	0.10133	1.96	0.058
degLC_CA − *degLC_UC*	0.15657	0.33362	5.63	0.000

Table 4 Paired samples test for *errFra* measure

	95% confidence interval of the difference		*t*	Sig. (2-tailed)
	Lower	Upper		
errFra_CA − *errFra_UC*	−0.92842	0.34018	−0.94	0.352

Table 5 Wilcoxon signed-rank test for *errAgg* measure

Items	Ranks	*N*	Mean rank	Sum of ranks
errAgg_CA − *errAgg_UC*	Negative ranks	20[a]	13.63	272.50
	Positive ranks	5[b]	10.50	52.50
	Ties	9[c]		
	Total	34		

[a] *errAgg_CA* < *errAgg_UC*

[b] *errAgg_CA* > *errAgg_UC*

[c] *errAgg_CA* = *errAgg_UC*

6.1.1 *Functional completeness w.r.t. a reference model*

By applying the Kolmogorov–Smirnov test, we noted that both *degFEC* and *degLCC* measures had a normal distribution ($p > 0.5$); the paired sample test was applied to verify the null hypotheses H_{10} and H_{20}.

As we can see in Table 3, there is a medium significance difference ($p = 0.05$) between the Use Cases and the Communication Analysis techniques, with respect to the degree of functional encapsulations completeness (*degFEC*). Besides, there is a very high significance difference ($p = 0.000$) with respect to the degree of linked communications completeness w.r.t. a reference model (*degLCC*). Therefore, H_{10} and H_{20} are refuted with a 95% confidence and the alternative hypotheses H_{11} H_{21} are corroborated. This means that Communication Analysis allows obtaining RS with greater degree of functional encapsulations and linked communications completeness than Use Cases.

This outcome can be explained by the fact that Communication Analysis methodological guidelines for the identification and specification of system functions follow a more systematic procedure than Use Cases guidelines. Also, Communication Analysis approaches functional requirements specification from a business process perspective; this way, temporal precedence relations of the specified processes facilitate the discovery of missing communicative events, thus contributing to higher completeness. Moreover, the Communicative Event Diagram technique devotes a modelling primitive to linked communications (outgoing communicative interactions) so they are explicitly specified.

Table 6 Test statistic-significance

	errAgg_CA − *errAgg_UC*
Z	−3.062
Asymp. Sig. (2-tailed)	0.002

6.1.2 *Appropriate granularity*

The Kolmogorov–Smirnov test was applied to normality test of both *errFra* and *errAgg* measures. As we note that only *errFra* measure had a normal distribution ($p > 0.5$), Paired Sample Test was also applied to verify the null hypothesis H_{30}. We note in Table 4 that, with respect to fragmentation errors (*errFra*), there is not a significant difference ($p = 0.352$) between Use Cases and Communication Analysis. Therefore, hypothesis H_{31} was not corroborated.

By using the Wilcoxon signed-rank non-parametric test for verifying null hypothesis H_{40}, we observe in Table 5 that 20 out of 34 subjects made a greater number of functional aggregation errors when applying Use Cases than when applying Communication Analysis. This statistical differences presented a high significance level (see Table 6, $p = 0.002$). Therefore, hypothesis H_{41} was corroborated with 95% confidence.

Applying Communication Analysis leads to functional requirements specifications with a more appropriate granularity than applying Use Cases (H_{41} was corroborated). This outcome may be explained by the fact that Communication Analysis methodological guidelines for Communicative

Table 7 Descriptive statistics for perceived ease of use

Statistics	Communication Analysis (PEOUA)	Use Cases (PEOUB)
N	22	25
Average	3.04	2.98
Standard dev.	0.69	0.72
Minimum	1.67	1.67
Maximum	4.17	4.33

Table 8 Wilcoxon signed-rank test for ease of use

Ranks	*N*	Mean rank	Sum of ranks
Negative rank	6[a]	13.42	80.50
Positive rank	13[b]	8.42	109.50
Ties	0[c]		
Total	19		

[a]PEOUB < PEOUA

[b]PEOUB > PEOUA

[c]PEOUB = PEOUA

Event Diagrams are based on more objective and prescriptive criteria that those of Use Case Diagrams.

6.2 Perceived efficacy of the RS methods

6.2.1 Perceived ease of use

First, the scores of each subject were averaged over the six questions that are relevant for determining the perceived ease of use (Q1, Q2, Q4, Q6, Q10, and Q12). Descriptive statistics were then calculated for both methods (see Table 7).

Note that the averages obtained are close to the value 3 (middle score on the Likert scale). Values ranged from a low of 1.67 to a maximum of 4.17 (A) and 4.33 (B). Furthermore, standard deviations of 0.69 (A) and 0.72 (B) were obtained, implying that the averages obtained for both techniques are representative.

In order to verify the null hypothesis H_{50}, the Wilcoxon signed-rank non-parametric test was applied, by comparing the averages of two samples of the respective two techniques to determine whether there are differences between them.

In Table 8, we observe that 13 out of 19 subjects perceived Use Cases as easer to use than Communication Analysis. However, we have not detected a significant level ($p = 0.559$). A possible interpretation for this outcome is that Use Cases guidelines are expressed in more informal terms than Communication Analysis guidelines; this fact makes them more understandable at first glance.

Table 9 Descriptive statistics for perceived usefulness

Statistics	Communication Analysis (PUA)	Use Cases (PUB)
N	22	25
Average	3.48	3.39
Standard dev.	0.39	0.55
Minimum	2.71	1.86
Maximum	4.17	4.80

Table 10 Wilcoxon signed-rank test for perceived usefulness

Ranks	*N*	Mean rank	Sum of ranks
Negative rank	10[a]	9.05	90.50
Positive rank	6[b]	7.58	45.50
Ties	3[c]		
Total	19		

[a]PUB < PUA

[b]PUB > PUA

[c]PUB = PUA

6.2.2 Perceived usefulness

We averaged the scores assigned by each subject over the seven relevant questions for determining perceived usefulness (Q3, Q5, Q7, Q8, Q9, Q13, and Q14). Descriptive statistics were then calculated for both RE methods (see Table 9). In order to determine whether significant differences exist between both methods with respect to perceived usefulness, null hypothesis, H_{60}, was verified by using the Wilcoxon signed-rank test.

As we can see in Table 10, 10 out of 19 subjects perceived Communication Analysis as more useful than Use Cases. This difference was statistically significant at medium level ($p = 0.024$). A possible interpretation for this outcome is the fact that the Communicative Event Diagram allows specifying more analytical information (e.g. communicative interactions and precedence relations, which are external aspects of the IS) than Use Case Diagrams (e.g. «include» relations are typically used for decomposition and reuse purposes, which are design-time decisions [14] and Jacobson even discourages their use [55]). Also, Communication Analysis guidelines, although perceived as less easy to understand, are more prescriptive and, therefore, may lead to more homogeneous specifications. Moreover, the unity criteria for communicative events are based on systems theory and communication theory in order to improve the adequacy of event granularity.

To clarify the previous point, we analyzed separately each one of the questions Q8, Q9 and Q13, which respectively relate to: (1) Distinction between external and inter-

Table 11 The Wilcoxon signed-rank test for Items Q8, Q9, and Q13 related to perceived usefulness

Items	Ranks	*N*	Mean rank
Distinction between external and internal interaction	Negative rank	5[a]	4.50
	Positive rank	4[b]	5.63
	Ties	9[c]	
Q8B–Q8A	Total	18	
Homogeneity of functional specification: Q9B–Q9A	Negative rank	4[d]	4.00
	Positive rank	2[e]	2.50
	Ties	11[f]	
	Total	17	
Adequate level of granularity: Q13B–Q13A	Negative rank	9[g]	7.00
	Positive rank	3[h]	5.00
	Ties	6[i]	
	Total	18	

[a] Q8B < Q8A
[b] Q8B > Q8A
[c] Q8B = Q8A
[d] Q9B < Q9A
[e] Q9B > Q9A
[f] Q9B = Q9A
[g] Q13B < Q13A
[h] Q13B > Q13A
[i] Q13B = Q13A

nal interaction; (2) Homogeneity of functional specification; (3) Adequate level of granularity of functional specification.

In Table 11, we note that 9 out of 18 subjects perceived that Communication Analysis facilitates specifying functional requirements with a more appropriate level of granularity than the Use Cases. Only 3 out of 18 subjects perceived the opposite. This statistically difference presented a medium significance level ($p = 0.04$).

The number of ties in for the other two questions does not allow to draw any conclusions about them.

7 Validity evaluation: threats

This section discusses issues with the potential to threaten the validity of the experiment [9].

7.1 Conclusion validity

We verified that the subjects had a homogeneous background by means of a questionnaire, so there is no threat due to random heterogeneity of subjects, which could give rise to greater variability in the measures. As a trade-off, homogeneity limits external validity.

The proposed measures related to actual efficacy have been theoretically reasoned and intend to be quite objective. However, an empirical testing of the metrics is advisable, in order to ensure the reliability of measures; this is planed as future work. Also, we plan to have the subjects' models reviewed by more expert reviewers and to perform agreement rounds; this will allow assessing the level of objectiveness of the metrics and the measuring procedure (inter-reviewer agreement).

With respect to the reliability of perception-based measures, we conducted a reliability analysis on the survey using the Chronbach alpha technique. The generic value obtained was 0.72, indicating that the items on the survey are adequately reliable. However, according to Garson [58], this score could be improved with a cut-off of 0.80 for a "good" scale.

7.2 Internal validity

Instrumentation is the effect caused by the instruments used in the experiment; in particular, the fact that paper form surveys are error-prone. To minimise this threat, the transcription of paper forms into spreadsheets and statistical analysis tools was double-checked by two experimenters. However, we could not avoid subjects making errors which reduced the number of valid observations (e.g. identification of the evaluated method, identification of subject, unanswered questions). Therefore, in the future, we plan to use software-based surveys that prevent errors while subjects fill out relevant data.

There is a risk related to the allocation of subjects to groups. Letting students decide which group to join according to their availability was a mistake; it led to imbalanced groups and the resulting groups cannot be assumed to have the very same characteristics (i.e. motivation). A lesson learned is that subjects should be allocated randomly.

We acknowledge a threat of maturation; that is, during the second round, the subjects already know the photography agency problem statement. We intended to also make a comparison of the results of the first round, but this was not possible due to the imbalance between both groups.

Table 12 Inter-item correlation analysis of the perception-based variables

	Perceived ease of use						Perceived usefulness									
	Q1	Q2	Q4	Q6	Q10	Q12	Q3	Q5	Q7	Q8	Q9	Q13	Q14	CV	DV	Valid
Q1	1.00	0.27	0.61	0.35	0.80	0.78	0.45	0.24	0.14	0.40	0.19	0.06	0.04	0.63	0.22	Yes
Q2	0.27	1.00	0.23	0.33	0.30	0.38	0.26	0.24	0.50	0.34	0.04	0.09	0.07	0.42	0.22	Yes
Q4	0.61	0.23	1.00	0.46	0.72	0.65	0.32	0.45	0.42	0.38	0.28	0.11	0.18	0.61	0.31	Yes
Q6	0.35	0.33	0.46	1.00	0.51	0.48	0.39	0.66	0.31	0.27	0.18	0.36	0.70	0.52	0.41	Yes
Q10	0.80	0.30	0.72	0.51	1.00	0.76	0.37	0.38	0.20	0.63	0.20	0.18	0.11	0.68	0.29	Yes
Q12	0.78	0.38	0.65	0.48	0.76	1.00	0.41	0.20	0.14	0.38	0.04	−0.02	0.14	0.67	0.18	Yes
Q3	0.45	0.26	0.32	0.39	0.37	0.41	1.00	0.31	0.26	0.30	−0.08	0.32	0.49	0.37	0.37	Yes
Q5	0.24	0.24	0.45	0.66	0.38	0.20	0.31	1.00	0.44	0.49	0.32	0.35	0.51	0.49	0.36	Yes
Q7	0.14	0.50	0.42	0.31	0.20	0.14	0.26	0.44	1.00	0.22	0.36	0.50	0.26	0.44	0.29	Yes
Q8	0.40	0.34	0.38	0.27	0.63	0.38	0.30	0.49	0.22	1.00	0.19	0.48	0.07	0.39	0.40	Yes
Q9	0.19	0.04	0.28	0.18	0.20	0.04	−0.08	0.32	0.36	0.19	1.00	0.29	0.15	0.32	0.15	Yes
Q13	0.06	0.09	0.11	0.36	0.18	−0.02	0.32	0.35	0.50	0.48	0.29	1.00	0.47	0.48	0.13	Yes
Q14	0.04	0.07	0.18	0.70	0.11	0.14	0.49	0.51	0.26	0.07	0.15	0.47	1.00	0.42	0.21	Yes

7.3 Construct validity

Two experimenters are authors of Communication Analysis. In order to reduce the threat of bias, two experimenters without expectancies have been involved.

A reference model of low quality is a threat. However, the authors have been using the photography agency case for research and education for more than 10 years, and its conceptual model is highly agreed by now. A three-person expert modelling committee made the final adjustments.

The experiment includes a single IS description so it may under-represent the construct of all ISs.

Since the subjects were trained in both methods, the results of the second round may be affected by their previous knowledge. Again, the imbalance between groups did not allow a comparison of the first round.

In order to demonstrate that we have evidence for construct validity, an inter-item correlation analysis of the perception-based variables (PEOU, PU) was applied. To do so we used two criteria: Convergent Validity (CV), which refers to the convergence among different indicators used to measure a particular construct, and Discriminant Validity (DV), which refers to the divergence of indicators used to measure different constructs. Average DV should be lower than the average CV. The results of the validity analysis for each construct show that the CV value was higher than the DV value for all PEOU and PU items (see Table 12).

7.4 External validity

With respect to the use of students as experimental subjects, several works suggest that, to a great extent, the results can be generalised to industry practitioners [59]. In any case, we are aware that more experiments with a larger number of subjects are necessary.

We thoughtfully selected a representative problem statement. However, more empirical studies with other requirements specifications are necessary.

8 Conclusions and further work

Empirical evaluation of RE methods is a strong need in the area of requirements engineering. This paper adopts a theoretical framework proposed by Moody for method evaluation and comparison [24]. The framework is extended by refining quality goals and proposing metrics that allow their operationalisation. The focus is put on actual efficacy and four variables related to semantic completeness and granularity errors are defined. With respect to perceived efficacy, a questionnaire is used.

With regard to semantic completeness, most of the previous works propose a rating based on judgement (e.g. using a Likert scale). We propose measuring the degree of functional encapsulations completeness (*degFEC*) and the degree of linked communications completeness (*degLCC*) with respect to a reference model (which is agreed by an expert modelling committee). Also, we propose assessing whether a functional RS has an appropriate granularity with respect to a given set of unity criteria; this allows determining the number of functional fragmentation errors (*errFra*) and functional aggregation errors (*errAgg*).

We acknowledge that the proposed quality goals and metrics related to feasible functional completeness are more practical in experimental settings than in industrial practice (where a reference model of the domain is very unlikely to

be available). In any case, they provide a valuable strategy to feedback method designers.

A laboratory experiment has been carried out to compare two RE methods; namely Use Cases [10] and Communication Analysis [11]. Functional requirements specifications have been quantitatively evaluated with respect to the proposed metrics and qualitatively evaluated with respect to the MEM perception variables.

The following hypotheses were verified: Communication Analysis allows obtaining RSs with greater degree of functional encapsulation completeness and greater degree of linked communication completeness than Use Cases; also, Communication Analysis allows obtaining RS with less functional aggregation errors than Use Cases. With regards to functional fragmentation errors, the difference, although favourable for Communication Analysis, was not statistically significant. These outcomes can be due to the fact that Communication Analysis offers prescriptive methodological guidance to modularisation. We believe that use case-based methods would benefit from taking into account the unity criteria proposed by Communication Analysis [23].

In addition, our findings were also that Communication Analysis was perceived as more useful than Use Cases; the subjects perceived that it facilitates determining the appropriate level of granularity of functional specifications. Other characteristics of usefulness, namely homogeneity of functional specifications and the distinction between external and internal interactions were not empirically corroborated due to the low level of significance obtained. We also noted that Use Cases were perceived as easier to use than Communication Analysis (13 out of 19 subjects); however, this difference would have to be confirmed with an improved level of significance.

Empirical evaluations such as this experiment allow comparing methods and highlighting their strengths and weaknesses. However, theoretical evaluations allow understanding better why these differences arise. This is planned for future research.

We acknowledge that a higher number of subjects is needed to reconfirm these initial results. In addition, we plan to take into account experience in, or knowledge of, the use of RE methods as a relevant factor in further experiments. The experience collected in this first study will facilitate us to address several of the identified threats. We want to analyze in depth the "whys" of the obtained results. To do this, we plan carry to out an evaluation of actual effectiveness with respect to the other quality attributes and to analyze the causality relations between actual efficacy and perceived efficacy.

Acknowledgements Research supported by the Spanish Ministry of Science and Innovation (MICINN) project SESAMO (TIN2007-62894), Generalitat Valenciana project ORCA (PROMETEO/2009/015), the MICINN FPU grant (AP2006-02323), and co-financed by FEDER.

We are also grateful to Dr. Lefteris Angelis from Aristotle University of Thessaloniki (Greece) for his helpful advice on statistical analysis.

References

1. España S, Condori-Fernández N, González A, Pastor Ó (2009) Evaluating the completeness and granularity of functional requirements specifications: a controlled experiment. In: 17th IEEE international requirements engineering conference (RE'09), Atlanta, GA, USA. IEEE, New York, pp 161–170
2. Boehm B, McClean RK, Ufrig DB (1975) Some experience with automated aids to the design of large-scale reliable software. IEEE Trans Softw Eng 1(1):125–133
3. Wieringa RJ, Heerkens JMG (2004) Evaluating the structure of research papers: a case study. In: 2nd international workshop in comparative evaluation of requirements engineering, Kyoto, Japan. IEEE Press, New York, pp 41–50
4. Wieringa RJ, Heerkens JMG (2006) The methodological soundness of requirements engineering papers: a conceptual framework and two case studies. Requir Eng 11(4):295–307
5. Höfer A, Tichy WF (2007) Status of empirical research in software engineering. In: Empirical software engineering issues. Critical assessment and future directions. LNCS, vol 4336. Springer, Berlin, pp 10–19
6. Iivari J, Kerola P (1983) A sociocybernetic framework for the feature analysis of information systems development methodologies. In: Olle TW, Sol HG, Tully CJ (eds) Information systems methodologies: a feature analysis. North-Holland, Amsterdam, pp 87–139
7. Dieste O, Lopez M, Ramos F (2008) Updating a systematic review about selection of software requirements specification techniques. In: 11th workshop on requirements engineering (WER 2008), Barcelona, Spain
8. Siau K, Rossi M (1998) Evaluation of information modeling methods—a review. In: 31st Hawaii international conference on systems science (HICSS 1998), Kohala Coast, USA, vol 5. IEEE, New York, pp 314–322
9. Wohlin C, Runeson P, Höst M, Ohlsson MC, Regnell B, Wesslén A (2000) Experimentation in Software Engineering: an introduction. Kluwer, Dordrecht
10. Cockburn A (2000) Writing effective use cases. Addison-Wesley, Reading
11. España S, González A, Pastor Ó (2009) Communication Analysis: a requirements elicitation approach for information systems. In: 21st international conference on advanced information systems, Amsterdam, The Netherlands. LNCS. Springer, Berlin
12. Wieringa RJ (1996) Requirements engineering: frameworks for understanding. Wiley, New York
13. Dobing B, Parsons J (2006) How UML is used. Commun ACM 49(5):109–113
14. Simons AJH (1999) Use cases considered harmful. In: Technology of object-oriented languages and systems, Nancy, France. IEEE Computer Society, Los Alamitos, pp 194–203
15. Dano B, Briand H, Barbier F (1997) A use case driven requirements engineering process. In: Third IEEE international symposium on requirements engineering (RE'97), Antapolis, MD. IEEE, New York
16. Rolland C, Achour CB (1998) Guiding the construction of textual use case specifications. Data Knowl Eng J 25(1–2):125–160
17. Constantine LL, Lockwood LAD (1999) Software for use: a practical guide to the models and methods of usage-centered design. Addison-Wesley, Reading
18. Cockburn A (1997) Structuring use cases with goals. J Object-Oriented Program

19. Phalp KT, Jonathan V, Cox K (2007) Improving the quality of use case descriptions: empirical assessment of writing guidelines. Softw Qual Control 15(4):383–399
20. Cox K, Phalp K, Shepperd M (2001) Comparing use case writing guidelines. In: 7th international workshop on requirements engineering: foundations for software quality (REFSQ'2001), Interlaken, Switzerland
21. Pastor O, González A, España S (2007) Conceptual alignment of software production methods. In: Conceptual modelling in information systems engineering. Springer, Heidelberg, pp 209–228
22. González A, España S, Pastor O (2008) Towards a communicational perspective for enterprise information systems modelling. In: IFIP WG 8.1 working conference on the practice of enterprise modeling, Stockholm, Sweden. LNBIP. Springer, Berlin
23. González A, España S, Pastor O (2009) Unity criteria for business process modelling: a theoretical argumentation for a software engineering recurrent problem. In: 3rd international conference on research challenges in information science (RCIS 2009), Fes, Morocco. IEEE, New York
24. Moody DL (2003) The Method Evaluation Model: a theoretical model for validating information systems design methods. In: Proceedings of the 11th European conference on information systems (ECIS 2003), Naples, Italy, 16–21 June 2003
25. Yadav SB, Bravoco RR, Chatfield AT, Rajkumar TM (1988) Comparison of analysis techniques for information requirement determination. Commun ACM 31(9):1090–1097
26. Lindland OI, Sindre G, Sølvberg A (1994) Understanding quality in conceptual modeling. IEEE Softw 11(2):42–49
27. Moody DL, Sindre G, Brasethvik T, Solvberg A (2003) Evaluating the quality of information models: empirical testing of a conceptual model quality framework. In: 25th international conference on software engineering (ICSE 2003), Portland, USA, pp 295–305
28. Moody DL, Sindre G, Brasethvik T, Solvberg A (2002) Evaluating the quality of process models: empirical testing of a quality framework. In: 21st international conference on conceptual modeling. Springer, Berlin
29. Larman C (1997) Applying UML and patterns. Prentice Hall, New York (see p 53)
30. Övergaard G, Palmkvist K (2004) Use Cases: patterns and blueprints. Addison-Wesley, Reading (see Part V)
31. Kulak D, Guiney E (2000) Use cases: requirements in context. Addison-Wesley, Reading, pp 94–95
32. Langlois RN (2002) Modularity in technology and organization. J Econ Behav Organ 49(1):19–37
33. Reijers H, Mendling J (2008) Modularity in process models: review and effects. In: 6th international conference on business process management (BPM 2008), Milan, Italy. LNCS, vol 5240. Springer, Berlin, pp 20–35
34. Rescher N (1977) Methodological pragmatism: systems-theoretic approach to the theory of knowledge. Basil Blackwell, Oxford
35. Davis FD, Bagozzi RP, Warshaw PR (1989) User acceptance of computer technology: a comparison of two theoretical models. Manag Sci 35(8):982–1003
36. Davis AM, Overmyer S, Jordan K, Caruso J, Dandashi F, Dinh A, Kincaid G, Ledeboer G, Reynolds P, Sitaram P, Ta A, Theofanos M (1993) Identifying and measuring quality in a software requirements specification. In: 1st international software metrics symposium, pp 141–152
37. Pohl K (1994) The three dimensions of requirements engineering: a framework and its applications. In: 5th international conference on advanced information systems engineering, Paris, France. Pergamon, New York
38. Krogstie J, Sindre G, Jorgensen H (2006) Process models representing knowledge for action: a revised quality framework. Eur J Inf Syst 15(1):91–102
39. Falkenberg E, Hesse W, Lindgreeen P, Nilsson B, Oei JLH, Rolland C, Stamper RK, VanAssche F, Verrijn-Stuart A, Voss K (1998) FRISCO. A framework of information systems concepts. IFIP WG 8.1 report
40. Krogstie J, Lindland OI, Sindre G (1995) Towards a deeper understanding of quality in requirements engineering. In: 7th international conference on advanced information systems engineering. Springer, Berlin, pp 82–95
41. Moody DL, Shanks GG (1994) What makes a good data model? Evaluating the quality of entity relationship models. In: 13th international conference on the entity-relationship approach, Manchester, UK. Springer, Berlin, pp 94–111
42. Schuette R, Rotthowe T (1998) The guidelines of modeling—an approach to enhance the quality in information models. In: Conceptual modeling (ER'98), Singapore. LNCS, vol 1507. Springer, Berlin, pp 240–254
43. Schuette R (1999) Architectures for evaluating the quality of information models—a meta and an object level comparison. In: 18th international conference on conceptual modeling (RE 1999). Springer, Berlin, pp 490–505
44. Shanks GG, Darke P (1997) Quality in conceptual modelling: linking theory and practice. Asia-Pacific conference on information systems (PACIS 1007), Brisbane, pp 805–814
45. Moody DL (2005) Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions. Data Knowl Eng 55(3):243–276
46. Lockemann PC, Mayr HC (1986) Information system design: techniques and software support. Information processing, vol 86. North-Holland, Amsterdam
47. Zave P, Jackson M (1997) Four dark corners of requirements engineering. ACM Trans Softw Eng Methodol 6(1):1–30
48. Jacobson I (2004) Use cases—yesterday, today, and tomorrow. Softw Syst Model 3(3):210–220
49. Kroll P (2004) Dr Process how many use cases should you have in a system? IBM Rational developerWorks documentation. Accessed 02-2009
50. Juristo N, Moreno A (2001) Basics of software engineering experimentation. Kluwer, Boston
51. Achour CB, Rolland C, Maiden NAM, Souveyet C (1999) Guiding use case authoring: results of an empirical study. In: IEEE symposium on requirements engineering. IEEE, Los Alamitos
52. Cox K, Phalp K (2000) Replicating the CREWS use case authoring guidelines experiment. Empir Softw Eng 5(3):245–267
53. Anda B, Sjøberg DIK, Jørgensen M (2001) Quality and understandability of use case models. In: 15th European conference on object-oriented programming (ECOOP 2001). LNCS, vol 2072. Springer, Berlin, pp 402–428
54. Fortuna M, Werner C, Borges M (2007) Um modelo integrado de requisitos com casos de uso. In: Workshop de ingeniería de requisitos y ambientes software IDEAS 2007
55. Jacobson I (1987) Object-oriented development in an industrial environment. In: Conference on object-oriented programming systems, languages and applications, Orlando, FL, USA. ACM, New York, pp 183–191
56. OMG (2009) OMG unified modeling language (OMG UML), superstructure, V2.2. http://www.omg.org/cgi-bin/doc?formal/09-02-02. Accessed 02-2010
57. Basili V, Rombach HD (1988) The TAME project: towards improvement-oriented software environments. IEEE Trans Softw Eng 14(6):758–773
58. Garson D (1998) Scales and standard measures. North Carolina State University. Updated September 2008. http://www2.chass.ncsu.edu/garson/pa765/standard.htm
59. Runeson P (2003) Using students as experiment subjects—an analysis on graduate and freshmen student data. In: 7th international conference on empirical assessment in software engineering, pp 95–102