Analyzing the Understandability of Requirements Engineering Languages for CSCW Systems: A Family of Experiments

Miguel A. Teruel
LoUISE Research Group, Computing Systems Department, University of Castilla- La Mancha
Avda. España s/n, 02071 Albacete (Spain)
miguel@dsi.uclm.es

Elena Navarro1
LoUISE Research Group, Computing Systems Department, University of Castilla- La Mancha
Avda. España s/n, 02071 Albacete (Spain)
tel +34 967 599 200 ext. 2624, fax +34 967 599 343
enavarro@dsi.uclm.es

Víctor López-Jaquero
LoUISE Research Group, Computing Systems Department, University of Castilla- La Mancha, Avda. España s/n, 02071 Albacete (Spain)
victor@dsi.uclm.es

Francisco Montero
LoUISE Research Group, Computing Systems Department, University of Castilla- La Mancha
Avda. España s/n, 02071 Albacete (Spain)
fmontoro@dsi.uclm.es

Javier Jaen
ISSI Group, Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia
Avda. España s/n. 46022 Valencia (Spain)
fjaen@dsic.upv.es

Pascual González
LoUISE Research Group, Computing Systems Department, University of Castilla- La Mancha
Avda. España s/n, 02071 Albacete (Spain)
pgonzalez@dsi.uclm.es

1 Corresponding author
Analyzing the Understandability of Requirements Engineering Languages for CSCW Systems: A Family of Experiments

Abstract.

Context: A collaborative system is a special kind of software whose users can perform collaboration, communication and collaboration tasks. These systems usually have a high number of non-functional requirements, resulting from the users’ need of being aware of other users with whom to collaborate, that is, the workspace awareness.

Objective: This paper aims at evaluating two Requirement Engineering languages i* and CSRML (an extension of i*) in order to determine which is the most suitable one to specify requirements of collaborative systems, taking into account their special characteristics regarding collaboration and awareness.

Method: We performed a family of experiments comprising an original experiment and two replicas. They were performed by 30, 45 and 9 Computer Science students, respectively, from Spain and Argentina. These subjects filled in two understandability questionnaires once they analyzed the requirements models of two systems: an e-learning collaborative system and a conference review system with some collaborative aspects support. Both models were specified by using the evaluated languages.

Results: The statistical analysis of the family of experiments showed that the understandability was higher for the models specified with CSRML than for those specified with i*, especially for collaborative aspects. This result was also confirmed by the meta-analysis conducted.

Conclusions: CSRML surpasses i* when modeling collaborative systems requirements models due to the specific expressiveness introduced to represent collaboration between users and awareness and the new resorts to manage actors and roles.

Highlights.

> Two Requirements Engineering languages, namely i* and CSRML, were analyzed attending to understandability criteria of requirements for CSCW systems.
> A family of experiments at three different universities was conducted.
> The participants tried to understand the requirements models of two different systems.
> The obtained results were summarized by means of meta-analysis techniques.
> Results show that CSRML is more understandable than i* for dealing with requirements of CSCW systems.

Keywords:
Collaborative systems, awareness, controlled experiment, understandability, requirements engineering, meta-analysis

1 Introduction

Internet is continuously evolving not only in the way it is developed, but also in the way it is browsed and exploited. Internet sites have evolved from simple static sites created by a single developer, which just enabled visitors to read their contents, to complex dynamic sites, which can be created and modified by users in a collaborative way. It is worth taking a look at the top 100 most visited pages in 2011 [1] and see that, except for search engines and news websites, almost all of the top 100 ones are collaborative webs. In addition, the just mentioned search engines and news web sites are
becoming also collaborative by including new social features, such as the possibility of adding comments or recommendations to news or search results.

Nowadays, even classic applications like text processors are collaborative. For instance, Google Docs [2] enables several users to edit a text document simultaneously. These collaborative text processors are a good example of CSCW (Computer Supported Cooperative Work) systems [3], which are systems whose users can perform collaboration, communication and coordination tasks (3C). Collaborative systems, in a similar way to classical single-user systems, have to be specified by means of a set of requirements, whose accuracy and suitability are key to achieve the quality of the developed system. The main difference between the requirements of single-user systems and CSCW systems is the highly non-functional nature of the latter, because of the users’ need of being aware of the presence of other users with whom to perform the above mentioned 3C tasks, that is, the Workspace Awareness (WA).

WA was defined by Gutwin as “the up-to-the-moment understanding of another person’s interaction within a shared workspace” [4]. This WA involves knowledge about where others are working, what they are doing now, when a user performed some action, and what they are going to do next. Table 1 and Table 2 show the WA elements related to the present and to the past, respectively.

### Table 1. Elements of WA related to the present

<table>
<thead>
<tr>
<th>Category</th>
<th>Element</th>
<th>Specific questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who</td>
<td>Presence</td>
<td>Is anyone in the workspace?</td>
</tr>
<tr>
<td></td>
<td>Identity</td>
<td>Who is participating? Who is that?</td>
</tr>
<tr>
<td></td>
<td>Authorship</td>
<td>Who is doing that?</td>
</tr>
<tr>
<td>What</td>
<td>Action</td>
<td>What are they doing?</td>
</tr>
<tr>
<td></td>
<td>Intention</td>
<td>What goal is that action part of?</td>
</tr>
<tr>
<td></td>
<td>Artifact</td>
<td>What object are they working on?</td>
</tr>
<tr>
<td>Where</td>
<td>Location</td>
<td>Where are they working?</td>
</tr>
<tr>
<td></td>
<td>Gaze</td>
<td>Where are they looking?</td>
</tr>
<tr>
<td></td>
<td>View</td>
<td>Where can they see?</td>
</tr>
<tr>
<td></td>
<td>Reach</td>
<td>Where can they reach?</td>
</tr>
</tbody>
</table>

### Table 2. Elements of WA related to the past

<table>
<thead>
<tr>
<th>Category</th>
<th>Element</th>
<th>Specific questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>How</td>
<td>Action history</td>
<td>How did that operation happen?</td>
</tr>
<tr>
<td></td>
<td>Artifact history</td>
<td>How did this artifact come to be in this state?</td>
</tr>
<tr>
<td>When</td>
<td>Event history</td>
<td>When did that event happen?</td>
</tr>
<tr>
<td>Who</td>
<td>Presence history</td>
<td>Who was here, and when?</td>
</tr>
<tr>
<td>Where</td>
<td>Location history</td>
<td>Where has a person been?</td>
</tr>
<tr>
<td>What</td>
<td>Action history</td>
<td>What has a person been doing?</td>
</tr>
</tbody>
</table>

In order to deal with the specification of these special type of systems, we conducted an initial test, by using DESMET [5], to check which is the most adequate Requirements Engineering technique to model both awareness and quality requirements of CSCW systems [6]. The analyzed techniques were Use Cases [7,8], Goal-Oriented [9] and Viewpoints [10], and we concluded that Goal-Oriented techniques were the most promising ones for specifying this kind of requirements. Next, we conducted a second study comparing the suitability of three Goal-Oriented approaches for modeling CSCW
systems requirements [11]. The compared proposals were NFR Framework [12], i* Framework [13] and KAOS Methodology [9]. The analysis of the results showed that i* was the most suitable proposal.

Despite the previous studies that determined that i* was the most promising technique, we identified several issues when modeling collaborative systems with this language, such as the lack of expressiveness for awareness requirements, imprecise 3C task representation, and a not fully suitable management of roles and actors. Due to these problems, the original i* language was extended to try to solve the above mentioned deficiencies by creating CSRML (Collaborative Systems Requirements Modeling Language) [14].

As Basili et al. [15] claim, families of experiments can be used to draw relevant conclusions that an experiment alone cannot provide. For this reason, in this paper, a family of experiments is presented whose general goal is to test which language, CSRML or i*, has a better understandability when modeling requirements of CSCW systems. This family of experiments has been conducted by following the Wholin et al.’s guidelines [16] and it has consisted of an experiment and two replicas, all of them carried out by students of Computer Science from three different universities. These students were asked to fill in two understandability questionnaires once they had analyzed the requirements models of an e-learning collaborative activity and a conference review system with collaborative aspects support. Both requirements models were specified with the tested languages.

As aforementioned, understandability was the criteria analyzed in the family of experiments. Understandability has been defined by the International Organization for Standardization (ISO) as “the capability of the software product to enable the user to understand whether the software is suitable, and how it can be used for particular tasks and conditions of use” [17]. Indeed, considering the requirements specification as one of the products of the software process, it becomes critical to provide a specification understandable by every stakeholder to develop a suitable software. This fact is also supported by the IEEE830-1998 standard for requirements specifications [18], since it considers an understandable specification as one of the means to decrease the development effort by reducing later redesign, recoding, and retesting. Moreover, understandability has been also considered as a quality attribute in several quality standards. For instance, it is one of the five quality characteristics that an engineering model must have along with abstraction, accuracy, predictiveness and inexpensiveness [19]. Furthermore, ISO includes understandability as one of the six sub-characteristics of the usability quality factor. More precisely, usability is defined as “the capability of the software product to be understood, learned, used and attractive to the user, when used under specified conditions” [17].

Based on these considerations, this paper is structured as follows: Section 2 presents several works related to CSCW and Empirical Software Engineering, focusing on the evaluation of Requirements Engineering models. Next, in Section 3 i* and CSRML are briefly introduced. In section 4 the family
of experiments is presented by describing how it has been carried out and its main results. Finally, some conclusions and future work are drawn in Section 5.

2 Related Works

In traditional software applications, only one user is able to interact with the application. Therefore, there is no support for collaboration among users. Nevertheless, in the late eighties, Computer Supported Cooperative Work (CSCW, [3,20]) emerged as a new discipline to study collaborative systems, a special kind of software in which users can collaborate with each other by sharing resources, chatting, etc.

One of the key elements for the construction of CSCW systems is the proper application of a methodology that not only guides their development, but also takes into account the peculiarities that make them different from other systems. An example is AMENITIES [21], a methodology presented by Garrido, for the development of cooperative systems based on behavior models and tasks. This methodology pursues the development of a system model from a requirements model by following a straightforward method that allows stakeholders to represent tasks and relate them to both the users responsible for their accomplishment (by means of roles) and other relevant concepts of the problem’s domain (artifacts, information…). Despite being a complete development methodology, AMENITIES has an important lack of expressiveness for the representation of the Workspace Awareness (WA), a cornerstone for the development of this kind of systems, as aforementioned in Section 1. This is due to the use of a classical RE technique, Use Cases, that makes it difficult to represent non-functional requirements related to the WA [6].

Based on the aforementioned methodology, Molina et al. presented CIAM [22], a methodological proposal for the development of user interfaces for CSCW systems. This methodology improves AMENITIES by combining interaction, collaboration and information sharing aspects and supporting WA in a better way. Nevertheless, it also uses the same RE technique than AMENITIES. Therefore, CSCW requirements cannot be properly specified. Also based on Garrido’s proposal, Penichet et al. presented TOUCHE [23], a process model for the development of user interfaces for CSCW systems, which is user-centered and driven by tasks. This methodology does improve the requirements specification by enriching the Use Case specification with information regarding the collaboration among users. Because awareness requirements can be considered as non-functional requirements [6], this proposal adds a perception needs textual slot to the Use Cases specification to answer Gutwin’s questions about WA (listed in Table 1 and Table 2), instead of using a more non-functional oriented technique such as the Goal-Oriented ones discussed in [11].

Some interesting works can be found regarding WA. For instance, Dourish and Belloutti [24] identified awareness as a critical issue that facilitates the group progress by providing users with the shared feedback. Thanks to this awareness, cost of information production can be reduced. In addition,
users can both obtain more easily information from other participants and find relevant information about shared objects. Another interesting work about WA is presented by Gutwin and Greenberg in [4]. In this work, as it was introduced in the previous section, they presented a theoretical framework where WA elements are classified according to when an awareness event happens (present, past) and which its category is (who, what, where, how, when). In addition, several example questions are presented to help one to identify WA elements. Despite describing WA in a clear-cut manner, these works provide no guidelines about how to specify these elements in a requirements model. Therefore, it is necessary to create, or extend, an existing RE technique to provide the expressiveness needed to specify the awareness requirements.

In order to validate both existing and emerging proposals, Empirical Software Engineering (ESE, [16,25]) is currently getting more and more attention from the Software Engineering community. In fact, it has been applied to several processes, methods and tools for stages of the development process such as design [26], implementation [27] or testing [28]. For example, ESE has been used lately to evaluate UML diagrams comprehension [29], development methods [30] or software maintenance tasks [31].

The interest in ESE may also be noted in the area of Requirements Engineering (RE). Several works have been carried out to evaluate different RE techniques. For instance, Karlsson et al. [32] evaluated six methods to establish priorities of requirements specifications regarding some of their inherent characteristics by using objective and subjective measures. They found out that AHP (Analytical Hierarchy Process) was the most promising approach. In a more recent work [33], they performed an experiment to compare the previous technique with Planning Game (PG) by analyzing variables like time consumption, ease of use or accuracy. They finally concluded that PG is the most suitable for priority assignment. Another interesting work was carried out by Kamsties et al. to evaluate the understandability of black-box and white-box requirements models by means of a controlled experiment [34]. By means of their results, they established some guidelines to choose a requirements specification style depending on both the kind of project to be carried out and the phases to be done. In addition, Ricca et al. [35] carried out a family of experiments to assess whether the adoption of table-based acceptance tests affects the understandability of requirements. They demonstrated that the use of these tables has a considerable bearing on the understandability of the requirements regarding those proposals that use only textual descriptions.

Furthermore, there are empirical evaluations of techniques of software requirements inspections [36], techniques of negotiation for distributed software process [37], techniques of prioritization of requirements [33], assessments of how the introduction of new specification elements improve the requirements understandability [38], the introduction of metrics for evaluating the completeness and granularity of requirements [39], etc. As can be observed, all these proposals are related to different techniques that can be applied during the RE process.
However, the evaluation of languages for requirements specification has not received enough attention. There are some works focused on this issue, such as the one presented by Celko et al. [40] or Bente et al. [41], but as far as we know, only Al-subaie and Maibaum have carried out an empirical evaluation of Goal-Oriented based methods [42], and more specifically, of KAOS [9] and its tool, Objectiver [43]. They tried to detect possible problems while modeling KAOS models with Objectiver, such as the difficulties with completeness check and the traceability between requirements and their sources.

In a similar way to RE, ESE has been also applied in the area of CSCW systems [44,45]. For instance, Ali Babar et al. [46] presented an empirical study about the application of CSCW for the evaluation process of distributed Software Architectures (SA). They compared face-to-face (F2F) meetings with distributed meetings, concluding that distributed meetings for SA evaluation are as good as F2F meetings, and even better under some conditions. Another relevant work directly related to WA is the one presented by Gutwin and Greenberg [47]. In this study, they conducted a usability test of several user interfaces with different levels of WA, and they concluded that users were able to carry out the experimental task in a better way by using the highest level of WA. However, to the best of our knowledge, there has been no work previously conducted on the evaluation of RE techniques for the specification of CSCW requirements. This constitutes the main aim of this work.

3 Languages Overview

In this section the evaluated languages, namely CSRML and $i^*$, are presented in order to facilitate the comprehension of the family of experiments. In section 3.1, the $i^*$ language is briefly described and later, in Section 3.2 CSRML is introduced highlighting which concepts have been introduced to facilitate the modeling of requirements of CSCW systems.

3.1 $i^*$ (i-Star)

$i^*$ [13,48] is a language for modeling requirements of information systems. It distinguishes between concepts (see Figure 1) and relationships (see Figure 2). These concepts are:

![Figure 1. $i^*$ Concepts](image)
• An actor is a person or a system that has a relationship with the system-to-be. $i^*$ identifies three types of actors:

— Agent is an actor who has a physical representation, e.g. a person or a system.
— Role defines the behavior of an actor in a specific context. An actor can play several roles, and a role can be assigned to multiple actors.
— Position is a set of roles that be typically played by one agent. An agent can play several positions.

• Goal: A goal answers “why?” questions. It describes a certain state of the world that an actor would like to achieve. However, a goal does not prescribe how it should be achieved.

• Task: A task specifies a particular way of doing something. Usually, a task consists of a number of steps (or sub-tasks) that an actor must perform to execute it.

• Resource: A resource is a (physical or informational) entity that the actor requires to achieve a goal or perform a task. The main concern about a resource is whether it is available and from whom.

• Softgoal: A softgoal is a condition in the world that the actor would like to achieve, but unlike the concept of (hard) goal, the condition to achieve it is not sharply defined. A softgoal is typically a quality attribute that constrains other element, such as a goal, a task or a resource. A softgoal is considered to be fulfilled if there is sufficient positive evidence for its fulfillment and little evidence against it.

The previous concepts are related to each other by using this set of relationships (see Figure 2):

- **Dependency**: it is a relationship between a depender and a dependee for a dependum. The depender and the dependee are actors and the dependum can be a goal, a task, a resource, or a softgoal. The depender depends on the dependee for achieving a goal, performing a task, or using a resource. If
the dependee fails to provide the depender with the required dependum, it becomes difficult or impossible for the depender to achieve the goal, perform the task, or use the resource. Based on the type of dependum, * distinguishes four types of dependencies: (i) Goal dependency determines that the depender assumes that the dependee achieves the goal, but does not prescribe how it should achieve it; (ii) Task dependency defines that the dependee must perform the assigned task; (iii) Resource dependency determines that the depender depends on the availability of a physical or informational resource that is provided by the dependee; (iv) Softgoal dependency determines that the depender depends on the dependee to achieve a softgoal. The criteria to determine how to achieve the softgoal are not clearly defined. Typically, the dependee offers several alternatives for achieving the softgoal, and the judgment of whether the softgoal is achieved or not is up to the depender.

- **Means-end link**: A means-end link documents which softgoals, tasks, and/or resources contribute to achieve a goal. A means-end link also facilitates the documentation and evaluation of alternative ways to satisfy a goal, i.e., different decompositions of goals into subgoals, tasks, or resources.

- **Task decomposition link**: A task decomposition link describes the essential elements of a task. A task decomposition link relates the task to its components, which can be any combination of subgoals, sub-tasks, resources, or softgoals. The decomposition of a task can thus comprise sub-tasks that must be performed, sub-goals that must be achieved, resources that are needed, and softgoals that typically define quality goals for the task.

- **Contribution link**: A contribution link documents a positive (+) or negative (-) influence from tasks or softgoals to another softgoal. A contribution link describes whether a task or a softgoal contributes to satisfy a softgoal positively or negatively. It does not define precisely which kind of support is offered or the extent of the given support.

### 3.2 CSRML (Collaborative Systems Requirements Modeling Language)

In order to deal with the expressiveness needs for the specification of requirements of CSCW systems, and based on the two studies aforementioned [6,11], CSRML [14] has been created (Collaborative Systems Requirements Modelling Language). This language is an extension of * including some elements to model the specific collaboration features of CSCW systems. Next, the CSRML (Figure 3) concepts and relationships are described, except for those whose meaning is the same as in *:
- Role: it is a designator for a set of related tasks to be carried out. What makes i* roles different from CSRML ones is that the latter can be played by an actor that participates in individual or collaborative tasks (through participation links) and that can be the responsible for the accomplishment of a goal (through responsibility links). In addition, the graphical notation is also different from the i* role, as the concept of role/actor boundary is not used in CSRML anymore.

- Actor: it is a user, program, or entity with certain acquired capabilities (skills, category, and so forth) that can play a role while it executes (uses devices) or is responsible for actions [49]. An actor has to play a role (specified by means of a playing link, see Figure 3) in order to participate in the system.

- Task: The concept of task in CSRML is the same as in i*. They only differ in the notation introduced to specify the importance of a task: one, two or three exclamation marks, depending on the importance of the task. In addition, two types of tasks have been identified in CSRML:

  - Abstract task: it is an abstraction of a set of concrete tasks and, possibly, other elements. Participation links cannot be directly assigned to this kind of tasks, as it is meant to be used for task decomposition purposes.

  - Concrete task: These are the tasks the participants are involved in. The abstract tasks are refined into concrete tasks. Participants are assigned to concrete tasks through participation links. There are four types of concrete tasks:

    o Individual task is a task that an actor can perform without any kind of interaction with other actors.

    o Collaboration / Communication / Coordination tasks require two or more actors to be involved in order to perform any kind of collaboration / communication / coordination between them.
**Awareness softgoal**: CSRML specializes $i^*$ softgoal into awareness softgoal. It represents a specific need of perception of other user’s presence / actions, so that the task the user wants to perform would be affected negatively or even could not be done if it is not provided.

**Awareness resource**: This special type of resource refers to an implementation or a design solution to accomplish an awareness softgoal.

**Playing link**: A playing link is used to represent when an actor can play a role. This link has a guard condition that determines when a role can be played by an actor.

**Participation link**: A participation link denotes who is involved in a task. This link has an attribute to specify its cardinality, i.e., the number of users that can be involved in a task.

**Responsibility link**: A responsibility link assigns a role (played by an actor) to a (soft) goal or a task. This link represents who is the stakeholder responsible for a goal/task accomplishment. Nevertheless, if the role is responsible for a goal or task, this role is also responsible for the elements it is decomposed into, unless a responsibility link reaches one of its sub-elements. Although a stakeholder is responsible for the accomplishment of a task (or goal), it is not mandatory that the actor playing a role that represents this stakeholder participates directly in this task (or subtasks).

### 4 The Family of Experiments

In this section, we present the context (Section 4.1), design (Section 4.2), conduction procedure (Section 4.3) and the analyzed results (Section 4.4) of the family of experiments (Figure 4) performed in order to assess the understandability of CSRML language with regard to $i^*$. As the three members of the family have been designed in a similar way, only a single description will be provided for all of them, presenting the results for each one in section 4.4.

![Figure 4. Chronology of the family of experiments](image)

#### 4.1 Experimental Context

As aforementioned, in order to test the understandability of CSRML versus $i^*$ for modeling software requirements of CSCW systems, a family of controlled experiments has been carried out following the guidelines described in [25] and [16]. According to Bereiter [50], understanding “is a psychological process related to an abstract or physical object, such as a person, situation, or message whereby one is able to think about it and use concepts to deal adequately with that object”. Consequently, the main
aim of our family of experiments is the evaluation of the ability of the subjects to think and use concepts of CSRML/$i^*$ to properly deal with CSCW software requirements.

The main goal of this family of experiments is defined by using the GQM (Goal Question Metric) template [51] as follows: **analyze $i^*$ and CSRML requirements specification language for the purpose** of evaluating their understandability **from the viewpoint** of requirements engineering researchers **in the context of** undergraduate students. To address this goal, we defined our null hypothesis as shown in Table 3.

**Table 3. Main features of the family of experiments**

<table>
<thead>
<tr>
<th>Null-Hypothesis</th>
<th>$H_{0A}$: CSRML has the same average score for understandability as $i^<em>$ when modeling CSCW systems requirements. $H_{1A}$: $\neg H_{0A}$ $H_{0B}$: The average score for understandability is the same regardless the domain used in the experiment. $H_{1B}$: $\neg H_{0B}$ $H_{0AB}$: CSRML has the same average score for understandability as $i^</em>$ when modeling CSCW systems requirements, regardless the domain used in the experiment and viceversa. $H_{1AB}$: $\neg H_{0AB}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variable</td>
<td>Understandability of requirements modeling languages, measured by $Und$</td>
</tr>
<tr>
<td>Independent Variables</td>
<td>The domain to which the models were related and the language ($i^*$ and CSRML) used for modeling CSCW systems requirements</td>
</tr>
<tr>
<td>Location</td>
<td>University of Castilla – La Mancha (Albacete, Spain) Polytechnic University of Valencia (Valencia, Spain) National University of La Plata (La Plata, Argentina)</td>
</tr>
<tr>
<td>Date</td>
<td>May 2011 May 2011 June 2011</td>
</tr>
<tr>
<td>Subjects</td>
<td>30 Computer Science students (15 Group 1; 15 Group 2) 45 Computer Science students (23 Group 1; 22 Group 2) 9 PhD in Computer Science students (4 Group 1; 5 Group 2)</td>
</tr>
</tbody>
</table>

As shown in Table 3 the subjects were Computer Science students from both undergraduate and PhD levels. It must be highlighted that all of them had prior experience in Requirements Engineering, but no experience in $i^*$ or CSCW modeling. Moreover, there was no evidence about their previous knowledge regarding the domains used in the evaluation.

It is worth noting that although this work presents the results of a family of experiments, there is a previous background that has led to its development:

- In an initial work, we performed an analysis of three Requirements Engineering techniques (Use Cases, Viewpoints and Goal-Oriented) in order to assess which one is the most suitable to model requirements of CSCW systems, concluding that Goal-Oriented is the most promising technique for this task [6].
- Next, we analyzed three Goal Oriented approaches, NFR Framework, KAOS and $i^*$, to identify which one is the most suitable for modeling CSCW system requirements [11]. Despite none of them satisfied all the expressiveness needs of CSCW system requirements, $i^*$ was found as the most powerful one to specify this kind of systems.
- Although $i^*$ was selected as the best alternative to model CSCW requirements, we found several problems, like the lack of expressiveness to specify both the collaboration between users and the awareness concept, as well as to manage actors/roles in a proper way. To solve these expressiveness issues within the scope of CSCW systems, an extension to $i^*$ was made by adding a set of features related to collaborative task specification, resulting in the CSRML language [14].
Indeed, the main aim of this family of experiments is to test whether CSRML really accomplishes its objective of modeling CSCW requirements by properly addressing the issues found in the aforementioned works.

4.2 Experimental Design

The main goal of the family of experiments is to measure the variable understandability (Und) of CSRML requirements models versus $i^*$ ones. For this reason, the ability to understand the experimental material properly was assessed by using the number of correct answers divided by the total number of questions, being Computer Science students our experimental units, and more specifically, their test results. These students had previous knowledge about Software Engineering and Requirements Engineering. For this reason, these students were familiar with requirements modeling, but not with any Goal Oriented approach. Our dependent variable was the understandability of requirements modeling languages, measured by Und, and the independent variables were the domain which the diagrams were related to and the used language $i^*$ and CSRML. Then, the following null hypothesis was considered: CSRML has the same score for understandability as $i^*$ when modeling CSCW systems requirements ($H_{OA}$ in Table 3). Two additional hypotheses were also considered, $H_{OB}$ and $H_{GAB}$, which are also described in Table 3 related to the significance of the domain and the interaction language*domain, respectively. More details about the family of experiments can be found in Table 3.

Each conducted experiment consisted in understanding a piece of a requirements model, specified by using either $i^*$ or CSRML, of two different CSCW systems. The first system was an e-learning system [14]. Specifically, the jigsaw learning activity was selected, also known as Experts meeting. It is a cooperative-learning technique where students individually do some research about a proposed problem and then they teach each other what they have learnt by sharing their individual view of the problem. The second system was a collaborative conference review system. More concretely, the part of the system related to the reviewing process was selected. This process is carried out in collaboration between three reviewers (see Appendix 1). In order to test the understanding of the requirements specifications, the students had to fill in a questionnaire for each system consisting of true or false answers. The questions for each system were exactly the same, regardless the particular language used to specify it. These questions were focused on collaboration between users, with special emphasis on the awareness techniques. One of the requirements specifications model and its related questionnaire can be found in Appendix 1 (full material can be found at [52]). In addition, the time the students required to finish the test was also gathered for statistical purposes.

In each performed experiment, the subjects were distributed into two groups (G1 and G2) in order to assign the two different system specifications to each one. This distribution of the subjects into groups 1 and 2 was made so that two students belonging to the same group were not side by side.
After the division of students into groups, they received an introductory session (about 60 minutes), where the instructor introduced both languages and systems. These explanations were neutral regarding the independent variable (whether using i* or CSRML).

For each one of the above mentioned systems two models were specified by using i* and CSRML, respectively. It was decided that G1 performed the experiment first trying to understand the i* model of the jigsaw activity and then just after that, the CSRML model of the reviewing process. G2 made exactly the opposite, to understand the jigsaw activity modeled with CSRML, and later the reviewing process specified by using i*. In addition, additional experimental material was elaborated to provide the students with a detailed description of both i* and CSRML, as well as a brief introduction to the concept of workspace awareness. The experiment design is summarized in Table 4. This process of assigning subjects to 4 different treatments was derived by using a 2x2 factorial design with confounded interaction [53] for the two dependent variables (System and Language). Actually, within a system, the language variable changes together with the group of subjects. Therefore, the learning effect is cancelled out.

Table 4. Experiments design

<table>
<thead>
<tr>
<th>Domain</th>
<th>Jigsaw</th>
<th>Conference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>i*</td>
<td>Group 1</td>
</tr>
<tr>
<td></td>
<td>CSRML</td>
<td>Group 2</td>
</tr>
</tbody>
</table>

The subjects were filtered out. We decided that they should have some knowledge about Requirements Engineering, but not about Goal-Oriented Requirements Engineering. Finally, we tried to keep a consistent population age. Therefore, we decided that those subjects meeting at least one of the following criteria would be dropped out from the experiment:

- The student age exceeds in five years the class mean age.
- The student is not familiar with Requirements Engineering.
- The student has some previous experience with i* or other Goal-Oriented techniques.

It was also decided that if a student wanted to leave the experiment without finishing it, he/she would be interviewed about the cause and this fact would be noted in the experiment results. Also, any possible interruption of the experiment would be also recorded.

It is worth noting that both experimental design and analysis was supervised by an statistician, as suggested in [25], due to some possible issues in this family of experiments, such as the reduced number of subjects of one of the replicas. It must be also highlighted that the questionnaires were elaborated by means of several iterations thanks to the collaboration with the Software Engineers of Symbia IT Corp, a Spin-off company with a long background in the development of collaborative systems. First of all, these Software Engineers established the issues to be deal with, that is, the questions should be devoted to collaborative tasks, roles-actors, importance and awareness. Later on,
and according to these issues, the questions were defined and refined based on the two systems used to carry out the family of experiments.

4.3 Conducting the experiments

The experiment and replicas were performed in a single classroom, where the subjects were properly located in order to avoid any possible interaction between them. As aforementioned in the experimental design, the students received the corresponding introductory session about \( i^* \) and CSRML. Additionally, the instructors were professors not related to the CSRML development to avoid a potential bias in the explanation. Before providing the students with the models, they were asked to provide the following information (See Appendix 1):

- Years of experience in Requirements Engineering.
- Years of experience with Goal-Oriented approaches.
- Average mark.
- Age.
- Gender.

With the aim of increasing subject’s interest and motivation, the instructors explained them that their final mark would be increased by half of a point over ten of their final mark. They also advised them that their evaluation would not be related to their performance in the experiment so that their answers would be evaluated anonymously. Finally, they were also informed that there was no time limit to carry out the task in order to avoid a possible ceiling effect.

Only three students were dropped out from the replica R1: two students belonging to G1 group, because they did not have previous knowledge about Requirements Engineering, and another student belonging to G2 group, because he was familiar with Goal-Oriented approaches. Moreover, only two students asked some questions that were related to the specification of the jigsaw activity with \( i^* \) during the original experiment (E1). During the experiment (E1) and the replicas (R1 and R2), the time needed by each student to carry out the tasks was also collected. Table 4 summarizes these data by showing the maximum (Max), minimum (Min) and average (\( \mu \)) times needed, along with the number of students (N) for each group.

Table 5. Time spent (mm:ss) on experiment E1 and replicas R1 and R2

<table>
<thead>
<tr>
<th>Group - Language</th>
<th>E1</th>
<th>R1</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Max</td>
<td>Min</td>
</tr>
<tr>
<td>( G1 - i^* )</td>
<td>15</td>
<td>12:00</td>
<td>06:00</td>
</tr>
<tr>
<td>( G2 - CSRML )</td>
<td>15</td>
<td>12:20</td>
<td>07:00</td>
</tr>
<tr>
<td>( G1 - CSRML )</td>
<td>15</td>
<td>11:10</td>
<td>04:09</td>
</tr>
<tr>
<td>( G2 - i^* )</td>
<td>15</td>
<td>11:00</td>
<td>05:27</td>
</tr>
</tbody>
</table>
4.4 Results Analysis

In this section, the results of each experiment and its corresponding analysis are presented. First, the results of the original experiment and the methodology used to analyze them are described. Just after that, the replica’s results are depicted obviating the analysis methodology as it is identical to the one used for the original experiment.

4.4.1 Original Experiment (E1)

Once the experiment concluded, the subjects’ data and responses were collected into an IBM SPSS Statistics v19 sheet and they were analyzed to exclude any observation that was not complete, or was stated by a subject who satisfied any of the criteria described in Section 4.2, or had not written down some of the requested data. All the data were properly written, therefore no observation was discarded.

First, Table 6 illustrates that the subjects’ groups who were given the CSRML models obtained a better score (mean values of 0.793 and 0.820) than those who tried to understand the i* models (0.573 and 0.593). Despite the clear difference between both systems, the results for the understandability were very similar between them (0.683 for the Jigsaw system vs. 0.707 for the Conference system), that is, the system had no influence on the understandability results.

![Table 6. E1 descriptive statistics for Und](image)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Language</th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jigsaw</td>
<td>iStar</td>
<td>0.573</td>
<td>0.1163</td>
<td>15</td>
</tr>
<tr>
<td>CSRML</td>
<td>0.793</td>
<td></td>
<td>0.1335</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>0.683</td>
<td></td>
<td>0.1663</td>
<td>30</td>
</tr>
<tr>
<td>Conference</td>
<td>iStar</td>
<td>0.593</td>
<td>0.2154</td>
<td>15</td>
</tr>
<tr>
<td>CSRML</td>
<td>0.820</td>
<td></td>
<td>0.1521</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>0.707</td>
<td></td>
<td>0.2164</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>iStar</td>
<td>0.583</td>
<td>0.1704</td>
<td>30</td>
</tr>
<tr>
<td>CSRML</td>
<td>0.807</td>
<td></td>
<td>0.1413</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>0.695</td>
<td></td>
<td>0.1917</td>
<td>60</td>
</tr>
</tbody>
</table>

Table 7. E1 ANOVA results

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III SS</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>0.008</td>
<td>1</td>
<td>0.008</td>
<td>0.324</td>
<td>0.572</td>
</tr>
<tr>
<td>Language</td>
<td>0.748</td>
<td>1</td>
<td>0.748</td>
<td>29.672</td>
<td>1.182E-06</td>
</tr>
<tr>
<td>Domain * Language</td>
<td>1.6667E-04</td>
<td>1</td>
<td>1.6667E-04</td>
<td>0.007</td>
<td>0.935</td>
</tr>
<tr>
<td>Error</td>
<td>1,412</td>
<td>56</td>
<td>0.025</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>31,150</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>2,169</td>
<td>59</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Despite the previous results, an ANOVA test (Table 7) was also performed to reject the null hypothesis $H_{0A}$, which is the most appropriate test for exploring the results of a 2x2 factorial design with interaction confounded [53,54]. We obtained a p-value of 1.182E-06 for the independent variable Language. Then, the null hypothesis $H_{0A}$ could be rejected with $\alpha=0.05$, because it can be concluded that there is a statistically significant difference between the results obtained for CSRML and i*. 

-15-
Nevertheless, the obtained p-values were higher than $\alpha$ for \textit{Domain} and \textit{Domain * Language} (these factors are not statistically significant), so we are unable to reject both null hypothesis $H_{0B}$ and $H_{0AB}$. This difference between both results has been graphically depicted in Figure 5(a).

![Figure 5](image)

**Figure 5. Graphical results for E1 (a) summary of results (b) results of each question for the Jigsaw system (c) results of each question for the Conference Review system**

Next, once it has been shown that CSRML gets a better score, the factors that contribute to this advantage are analyzed by examining the results grouped by question. For this aim, the global questions results are shown in two graphs, Figure 5(b) and Figure 5(c), to describe the results for each system.

The first notorious result that can be observed is that CSRML gets higher marks than $i^*$ for the three first questions. It is especially relevant that, for the first question, CSRML gets a success ratio of 100%. These three initial questions were related to collaboration tasks, so it can be stated that the collaboration representation is a clear strength of CSRML, being this a cornerstone for the specification of CSCW systems.

Both languages and domains get the same results for question 4. This question was related to a task that can be carried out by an individual user. This can be easily represented with both languages.

For questions related to roles (questions 5 and 6), CSRML supersedes $i^*$ one more time, because of its mechanisms to represent when an actor can play a role, and its task’s cardinalities representation. The same happens to awareness questions (7, 8 and 9) due to the new awareness elements introduced in CSRML. From these results, it can be concluded that the new awareness elements added to CSRML help users to identify both, awareness softgoals and implementations / design solutions thanks to the clear-cut way of specifying softgoals and resources as awareness elements.

The last question (10) was related to the understanding of the level of importance of tasks. In this scope, both languages provided similar facilities to the subjects.

Nevertheless, better results are obtained for the last question (or the same in case of the Conference system) when using $i^*$. For this reason, in future experiments it will be analyzed the impact of determining the task’s importance by using an alternative representation.
4.4.2 **First Replica (R1)**

As in the original experiment, an initial analysis of the collected data was performed. In this case, two students, belonging to G1 group, were discarded because they did not have previous knowledge about Requirements Engineering and another student, belonging to G2 group, stated he was familiar with Goal-Oriented approaches. For this reason, 42 subjects carried out the R1 replica, 21 for each group.

Once again, the subjects understood the CSRML models better than the equivalent i* models (0.748 vs. 0.538 for jigsaw system and 0.767 vs. 0.529 for Conference Review system), regardless of the system given (see Table 8). Therefore, the system had no influence on the results one more time, as the total results show, 0.643 for the Jigsaw system vs. 0.648 for the Conference Review system.

**Table 8. R1 Descriptive statistics for Und**

<table>
<thead>
<tr>
<th>Domain</th>
<th>Language</th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jigsaw</td>
<td>iStar</td>
<td>0.538</td>
<td>0.1244</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>CSRML</td>
<td>0.748</td>
<td>0.1401</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>0.643</td>
<td>0.1684</td>
<td>42</td>
</tr>
<tr>
<td>Conference</td>
<td>iStar</td>
<td>0.529</td>
<td>0.1678</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>CSRML</td>
<td>0.767</td>
<td>0.1461</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>0.648</td>
<td>0.1966</td>
<td>42</td>
</tr>
<tr>
<td>Total</td>
<td>iStar</td>
<td>0.533</td>
<td>0.1459</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>CSRML</td>
<td>0.757</td>
<td>0.1417</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>0.645</td>
<td>0.1820</td>
<td>84</td>
</tr>
</tbody>
</table>

**Table 9. R1 ANOVA results**

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III SS</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>4.762E-04</td>
<td>1</td>
<td>4.762E-04</td>
<td>0.023</td>
<td>0.881</td>
</tr>
<tr>
<td>Language</td>
<td>1.052</td>
<td>1</td>
<td>1.052</td>
<td>49.752</td>
<td>5.598E-10</td>
</tr>
<tr>
<td>Domain * Language</td>
<td>0.004</td>
<td>1</td>
<td>0.004</td>
<td>0.203</td>
<td>0.654</td>
</tr>
<tr>
<td>Error</td>
<td>1.691</td>
<td>80</td>
<td>0.021</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>37.720</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>2.748</td>
<td>83</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As in the previous experiment, an ANOVA test was performed (see Table 9) in order to reject the null hypothesis \(H_{0a}\). This time, with an \(\alpha\) of 0.05, a p-value of 5.598E-10 was obtained for the independent variable *Language*. Then, with a p-value < \(\alpha\), it can be concluded that there is a statistically significant difference between the results obtained with CSRML and i* again. In this replica, as in the E1 experiment, *Domain* and *Domain * Language* p-values are higher than \(\alpha\), thus there is not statistical difference between domains, and there is no influence between domains and languages, so we cannot reject \(H_{00b}\) and \(H_{00AB}\) again. Figure 6 (a) shows this difference by using a box-plot.
Figure 6. Graphical results for R1 (a) summary of results (b) results of each question for the Jigsaw system (c) results of each question for the Conference Review system

Before carrying out the ANOVA test, an analysis of the individual questions was performed by using the graphs shown in Figure 6 (b) and Figure 6 (c). In these graphs, it can be observed that for the first six questions, related to collaboration among users (questions 1 and 2) and actor/roles management (questions from 3 to 6), CSRML surpasses $i^*$, except for the third question related to the Conference Review system. This result can be explained considering that the students understood the review as a collaborative process.

Regarding the awareness concept evaluated in questions 7 and 8, which are related to the implementation of awareness techniques, CSRML gets better results one more time, but to a lesser degree than the previous questions. The results for question 9, related to awareness softgoals, is inconclusive, because $i^*$ supersedes CSRML for the Jigsaw system and it does the opposite for the Conference Review system, having both a similar difference. This can be explained because both of them use exactly the same representation, and the same subjects did not understand this question neither for the Jigsaw system nor the Conference Review system. Regarding the last question, related to the understanding of the importance degree of a task, on the contrary to the E1experiment, CSRML gets better results this time. Because of this contradictory result, more experiments will have to be carried out to draw a conclusion about the representation of the importance degree of tasks.

4.4.3 Second Replica (R2)

In order to conduct a new experiment with a different kind of subjects, a second replica was performed with PhD students. It was carried out at the University of La Plata (Argentina), far away from the first two experiments, in order to involve subjects with a different educational background. Unfortunately, only 9 subjects could carry out this replica. Nevertheless, this experiment provided us a new viewpoint about the previous results.

In order to analyze the obtained data, firstly a filtering process was carried out to determine if any subject met any of the criteria described in section 4.2. Fortunately, no one had to be discarded. For the third time, CSRML obtained better results than $i^*$, with mean values of 0.700 vs. 0.475 for the Jigsaw system and 0.825 vs. 0.520 for the Conference Review system (see Table 10).
Table 10. R2 Descriptive statistics for Und

<table>
<thead>
<tr>
<th>Domain</th>
<th>Language</th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jigsaw</td>
<td>iStar</td>
<td>0.475</td>
<td>0.2217</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>CSRML</td>
<td>0.700</td>
<td>0.1000</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>0.600</td>
<td>0.1936</td>
<td>9</td>
</tr>
<tr>
<td>Conference</td>
<td>iStar</td>
<td>0.520</td>
<td>0.1789</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>CSRML</td>
<td>0.825</td>
<td>0.0957</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>0.656</td>
<td>0.2128</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>iStar</td>
<td>0.500</td>
<td>0.1871</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>CSRML</td>
<td>0.756</td>
<td>0.1130</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>0.628</td>
<td>0.1994</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 11. R2 ANOVA results

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III SS</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>0.032</td>
<td>1</td>
<td>0.032</td>
<td>1.311</td>
<td>0.271</td>
</tr>
<tr>
<td>Language</td>
<td>0.312</td>
<td>1</td>
<td>0.312</td>
<td>12.739</td>
<td>0.003</td>
</tr>
<tr>
<td>Domain * Language</td>
<td>0.007</td>
<td>1</td>
<td>0.007</td>
<td>0.290</td>
<td>0.599</td>
</tr>
<tr>
<td>Error</td>
<td>0.343</td>
<td>14</td>
<td>0.025</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7.770</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>0.676</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dependent Variable: Und. a. R Squared = 0.493 (Adjusted R Squared = 0.384)

In this replica, due to the reduced number of subjects, the Und results had to be analyzed to check if they had similar variance and followed a normal distribution for both languages before carrying out an ANOVA test. In order to check the variance, a Levene contrast was used [53] obtaining a p-value of 0.520 (α = 0.05). In order to test the normal distribution, a Shapiro-Wilks normality test [53] was performed and the p-values were 0.263 for i* results and 0.248 for CSRML results. Therefore, the two requirements to use ANOVA were satisfied.

With the previous conditions satisfied, the ANOVA test (see Table 11) was carried out obtaining a p-value of 0.003 for the Language independent variable, so that the null hypothesis $H_{0A}$ could be rejected, that is, the average values for understandability are significantly different. It is worth noting that the results of this experiment, on the contrary to the two previous ones, show a little difference between the Jigsaw system (total value 0.600) and the Conference Review system (total value 0.656). Nevertheless, as in the two previous experiments, for Domain and Domain * Language the obtained p-values were higher than $\alpha$ (0.05). Therefore, it endorses that we cannot reject the null hypothesis $H_{0B}$ because the domain has no influence in the understandability of this replica. Furthermore, there is no interaction effect between the language and the domain factors, preventing us from rejecting the null hypothesis $H_{0AB}$. Again, for the third time, CSRML widely surpassed i* (see graphical results in Figure 7(a)).
Figure 7. Graphical results for R2 (a) summary of results (b) results of each question for the Jigsaw system (c) results of each question for the Conference Review system

Figure 7(b) and Figure 7(c) show the results of each question for the Jigsaw system and the Conference Review system, respectively. As in the previous experiments, for the 6 first questions CSRML obtained better results, except for question number 5 of the Jigsaw questionnaire, related to guard conditions. Questions related to the implementation of awareness techniques provide results that are less conclusive than for the other two experiments. Although CSRML still keeps its advantage for the Conference Review system, the results are not so good for the Jigsaw system. This can be explained because, as evidenced by post-experimental informal conversations with the subjects, some students had not fully understood some of the modeling of the awareness characteristic in CSRML. This led us to the conclusion that additional work on CSRML has to be done to improve the specification of awareness. Finally, while the results of CSRML for question 9 (related to awareness softgoals) surpasses widely those of $i^*$, results for question 10 are similar to those obtained in experiment E1, that is, the $i^*$ results are slightly better. Therefore, it can be concluded that the students did not understand the proposed way to assign importance to tasks.

4.5 Summarizing Results: Meta-analysis

In this section the results of the original experiment and its replicas are summarized. With this aim, first an ANOVA test was performed considering the origin of students and, later a meta-analysis was conducted like those performed in typical ESE studies [55,56].

As aforementioned and considering the consulted statistician’s suggestion, an analysis similar to that applied to the individual experiments was performed that consisted in a preliminary descriptive statistics analysis (see Table 12) and a three factors ANOVA test (see Table 13).
Table 12. Summarized descriptive statistics for Und

<table>
<thead>
<tr>
<th>Domain</th>
<th>Language</th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jigsaw</td>
<td>iStar</td>
<td>0.545</td>
<td>0.1319</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>CSRML</td>
<td>0.759</td>
<td>0.1341</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>0.653</td>
<td>0.1704</td>
<td>81</td>
</tr>
<tr>
<td>Conference</td>
<td>iStar</td>
<td>0.551</td>
<td>0.1859</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>CSRML</td>
<td>0.793</td>
<td>0.1439</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>0.670</td>
<td>0.2052</td>
<td>81</td>
</tr>
<tr>
<td>Total</td>
<td>iStar</td>
<td>0.548</td>
<td>0.1606</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>CSRML</td>
<td>0.775</td>
<td>0.1392</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>0.662</td>
<td>0.1882</td>
<td>162</td>
</tr>
</tbody>
</table>

Table 13. Family of experiments ANOVA results

Dependent Variable: Und. a. R Squared = 0.396 (Adjusted R Squared = 0.351)

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III SS</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>0.016</td>
<td>1</td>
<td>0.016</td>
<td>0.720</td>
<td>0.398</td>
</tr>
<tr>
<td>Language</td>
<td>2.094</td>
<td>1</td>
<td>2.094</td>
<td>65.056</td>
<td>1.752E-17</td>
</tr>
<tr>
<td>Domain * Language</td>
<td>0.008</td>
<td>1</td>
<td>0.008</td>
<td>0.426</td>
<td>0.559</td>
</tr>
<tr>
<td>Error</td>
<td>3.589</td>
<td>158</td>
<td>0.023</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>76.640</td>
<td>162</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>5,703</td>
<td>161</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

By means of these results, and due to the obtained p-value of 1.752E-17 for Language factor, the null hypothesis H₀A can be rejected. Therefore, definitively there is a statistical significant difference between CSRML and i*. Additionally, as the obtained p-value (see Table 13) for Domain was higher than the established α (0,05), it can be stated that the systems used for the tests did not have influence over the understandability. This can be also stated for the combination of the two independent variables (Domain * Language in Table 13). A box-plot summarizing these results can be seen in Figure 8(a).

Figure 8. Summarized graphical results (a) summary of results (b) results of each question for the Jigsaw system (c) results of each question for the Conference Review system

In order to get some final conclusions about how the features of CSRML and i* influence characteristics such as collaboration between user, roles and actors management and awareness representation, an analysis was carried out similar to that performed for each experiment, analyzing the final results question by question (see Figure 8(b) and (c)). The most important results are that for questions related to collaboration and actors/roles management (1 to 6), CSRML always gets better...
scores (and sometimes with a very high difference) than $i^*$. The same happens for questions 7, 8 and 9, related to the awareness concept, but this time the difference is not so obvious. Nevertheless, for question 10, which is related to tasks importance specification, CSRML has no significant difference (only 5% advantage for CSRML) versus $i^*$. This confirms what was also inferred after analyzing the individual experiments. Despite being a positive result for CSRML, this fact could confirm our assumption; CSRML’s mechanism for establishing the importance of a task may be not fully understandable. Due to this conclusion, different ways to set the importance of a task must be considered and evaluated in future works.

Once the results were summarized by means of a three-factor ANOVA test and the individual questions were analyzed, a meta-analysis was carried out. It is a common practice in disciplines with a strong experimental tradition, such as Medicine, Psychology or Physics, and also, Software Engineering. According to Dieste et al. [57], several experimental studies can be aggregated (or combined) by means of meta-analysis (also known as quantitative synthesis or research synthesis). Aggregated results are more reliable (and potentially more generally applicable) than individual experimental results.

In order to perform this meta-analysis, the Biostat’s Comprehensive Meta-Analysis 2.2 tool [58] was used. First, the Hedges’ $g$ metric [59] had to be computed to standardize the measures of each study, because they were taken in different environments. This metric was used to estimate the Global Effect Size (Table 14) of the evaluated factor, understandability of Requirement Engineering languages for CSCW systems. In this case, it is relative to the within group standard deviations. Moreover, the effect size of each study was also classified according to the obtained Hedges’ $g$ as small, medium and large [60]. The results of this computation are presented in Table 14.

<table>
<thead>
<tr>
<th>Study</th>
<th>Domain</th>
<th>$i^*$ mean</th>
<th>$i^*$ sd</th>
<th>CSRML mean</th>
<th>CSRML sd</th>
<th>n</th>
<th>Hedges’ $g$</th>
<th>Std. Err.</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>Conf.</td>
<td>0.593</td>
<td>0.215</td>
<td>0.820</td>
<td>0.152</td>
<td>15</td>
<td>-1.185</td>
<td>0.387</td>
<td>Large</td>
</tr>
<tr>
<td>E1</td>
<td>Jigsaw</td>
<td>0.573</td>
<td>0.116</td>
<td>0.793</td>
<td>0.134</td>
<td>15</td>
<td>-1.710</td>
<td>0.418</td>
<td>Large</td>
</tr>
<tr>
<td>R1</td>
<td>Conf.</td>
<td>0.529</td>
<td>0.168</td>
<td>0.767</td>
<td>0.146</td>
<td>21</td>
<td>-1.484</td>
<td>0.343</td>
<td>Large</td>
</tr>
<tr>
<td>R1</td>
<td>Jigsaw</td>
<td>0.538</td>
<td>0.124</td>
<td>0.748</td>
<td>0.140</td>
<td>21</td>
<td>-1.555</td>
<td>0.347</td>
<td>Large</td>
</tr>
<tr>
<td>R2</td>
<td>Conf.</td>
<td>0.520</td>
<td>0.179</td>
<td>0.825</td>
<td>0.096</td>
<td>4</td>
<td>-1.819</td>
<td>0.734</td>
<td>Large</td>
</tr>
<tr>
<td>R2</td>
<td>Jigsaw</td>
<td>0.475</td>
<td>0.222</td>
<td>0.700</td>
<td>0.100</td>
<td>5</td>
<td>-1.222</td>
<td>0.662</td>
<td>Large</td>
</tr>
<tr>
<td>Global Effect Size</td>
<td></td>
<td>-1.481</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.173</td>
<td>Large</td>
</tr>
</tbody>
</table>

Next, the meta-analysis technique to be applied was selected by following Dieste et al.’s guidelines [57]. For this reason, first, a parametric alternative had to be selected. According to the tables presented in [57], and the variances/effect sizes of the family, two methods were found applicable: Weighted Mean Difference (WMD) [59] and Pondered Response Ratio (PRR) [61]. As PRR is not widely used in Software Engineering, WMD was finally used. Considering the results shown in Figure 9, it can be observed that the p-value was ~0.000 for the meta-analysis (row Overall/column p-value in Figure 9) and the WMD was -1.531 (row Overall/column Std dif in means in Figure 9), so that the null
hypothesis $H_0$ could be rejected. Therefore, it can be definitively stated that there is a significant difference between CSRML and $i^*$ for modeling CSCW requirements. Figure 9 illustrates the weight of each study by means of the size of the rectangles. Note that some studies do not contribute equally to the overall results, being more accurate the estimations of those studies with a larger sample size (represented by means of the size of the squares).
Figure 9. Understandability meta-analysis
4.6 Threats to the Validity of the Family of Experiments

Next, the family of experiments is analyzed according to the four types of threats to Validity proposed by Wohlin et al. [16].

4.6.1 Conclusion Validity

In this family of experiments, as can be observed in both individual and meta-analysis, the score obtained for the different ANOVA tests is statistically significant. Therefore, it supports rejecting the initial null hypothesis with a large degree of certainty.

4.6.2 Internal Validity

The number of subjects that participated in the two first experiments was large enough (=30), according to the central limit theory [62]. It is necessary to obtain a statistically significant set of data. In R2, despite the number of participants was lesser than 30, it was successfully checked that the data followed a normal distribution and the variances between both data series were similar. Therefore, the conditions were valid to carry out the ANOVA test. It is worth noting that a statistician supervised all our statistical analysis.

Moreover, in order to avoid a possible source of bias, the questionnaires given to the subjects were reviewed by several experts in collaborative systems development not related to the CSRML development, who work for Symbia IT Corp., a university spin-off from Albacete (Spain). They helped us to elaborate the questionnaire taking into account the main features of CSCW systems.

Furthermore, we only used an excerpt of the models of the two systems during the experiments to prevent the students from getting tired or bored. In addition, with the aim of motivating the students, they were given half a point on their final mark of their respective subjects.

4.6.3 Construct Validity

A measure based on correct answers divided by the total number of answers was used to quantify the understandability of the models. Nevertheless, as future work, other guidelines such as the one presented in CTML [63] will be used to use an approach more closely related to learning from a sociological point of view.

4.6.4 External Validity

The models presented in this experiment were relatively easy to understand for the students. This issue could be noted as students did not ask about the models during the experiments. These models are also realistic as they specify requirements for real-life systems.

Moreover, the subjects had enough maturity level to participate in the experiment due to their previous courses on Software Engineering and Requirements Engineering. Students were selected to
perform this family of experiments because the task to be performed did not require a high level of industrial experience so that, according to [15,64], these experimental subjects can be considered appropriate in a similar way to other ESE studies [55,56].

5 Conclusions and Further Work

This family of experiments has its foundations in two previous works [6,11] that helped us to identify which RE technique is more suitable to model CSCW systems requirements. Nevertheless, these previous experiments led us to find that this ilk of specifications suffer from an important lack of expressiveness for some CSCW characteristics, such as user collaboration or awareness representation. To address these shortcomings, CSRML was proposed [14], an extension of i* to model CSCW systems requirements.

In order to test the understandability of CSRML versus i*, a family of experiments was performed. The participants of the experiment were 84 Computer Science students -30 second year students from Albacete (Spain), 45 second year students from Valencia (Spain) and 9 Ph.D. students from La Plata (Argentina). These subjects filled in two understandability questionnaires once they analyzed the requirements models of a collaborative e-learning activity and a collaborative conference review system, both specified by using the evaluated languages.

The most important result shown in the analysis of these experiments is that CSRML improves the understandability of CSCW requirements model versus i*. Moreover, the domain of the models and the origin of the participants had no influence in the results obtained. Other important facts were detected while analyzing the students’ answers. For example, it was found that CSRML models are more understandable due to its collaboration representation features, the new added awareness elements and the roles management approach used. This fact is supported by the analysis of the results of every question of the questionnaires (except for the tenth question). Nevertheless, an important issue regarding the expressiveness for establishing the importance of a task was also found. This issue has to be solved in future versions of the language, as it was not fully understandable in its current form by the subjects of the experiment.

Several challenges emerge as future work. One of them is broadening the scope of the evaluation already performed of CSRML vs. i*. Although the use of students as subjects of experiments has been proved suitable [15,64], we are planning new experiments that involve Software Engineering practitioners as the subjects. In these experiments they will be asked to analyze and modify some given requirements models. These experiments will consist in the specification of a requirements model based on a given system textual specification. Also, as a further work, we are planning additional evaluations for other criteria, such as usefulness, correctness, completeness, efficiency or effectiveness, as in the experiments described in [55,65,66]. In addition, we are considering also the
use of CTML [63] instead ratios of correct answers divided by total answers with the aim of providing our evaluation with a more sociological point of view.

Acknowledgements

This work has been partially supported by the grant (PEII09-0054-9581) from the Junta de Comunidades de Castilla-La Mancha and by the grant (DESACO, TIN2008-06596-C02-01) from the Spanish Government. We would like to thank Leandro Antonelli (Software Engineering teacher at UNLP), Francisco Parreño (Statistics teacher at UCLM) and Oscar Dieste (Empirical Software Engineering expert at UPM) for their advices during the course of this study. We would also like to thank Symbia IT Corp. software engineers for their feedback to help us to improve the questionnaires.

References

Approaches to Software Engineering (ENASE’11), SciTePress, Beijing, China, 2011: pp. 131-142.


Appendix 1: Experimental Material - An Example of an Understanding Task (Test 1)²

The following CSRML diagram describes a jigsaw learning activity, a cooperative-learning technique where students individually do some research about a proposed topic and then they teach each other what they have learned by sharing their individual view of the topic.

WRITE DOWN STARTING TIME (HH: MM: SS)

Answer the following True/false statements:
1. The task “Designate experts coordinator” is a collaborative task 

2. The task “Expose at team meeting” is a collaborative task 

3. Two different roles are required to perform the task “Give permission to an expert for exposing” 

4. Two different roles are required to perform the task “Attend experts meeting” 

5. A “Student” actor can play the “Expert” role at anytime 

6. Only an “Expert” is required to “Make experts report” 

7. “Participant list with exposition order” is an implementation of “Be aware of expert group member” softgoal 

8. “Experts report” is an implementation of “Make experts report” 

9. “See the other users activity” is a softgoal related to awareness 

10. Designate experts coordinator” task is more important than “Expose at experts meeting “ task 

² The full experimental material can be downloaded from https://www.dsi.uclm.es/trep.php?codtrep=DIAB-11-06-1