




On Modelers Ability to Build a Visual Diagram from a User Story Set: A Goal-Oriented Approach

Yves Wautelet¹ , Mattijs Velghe¹, Samedi Heng², Stephan Poelmans¹,
and Manuel Kolp²

¹ KU Leuven, Leuven, Belgium

{yves.wautelet, stephan.poelmans}@kuleuven.be

² Université catholique de Louvain, Louvain-la-Neuve, Belgium

{samedi.heng, manuel.kolp}@uclouvain.be

Abstract. [Context and Motivation] User Stories (US) are often used as requirement representation artifacts within agile projects. Within US sets, the nature, granularity and inter-dependencies of the elements constituting each US is not or poorly represented. To deal with these drawbacks, previous research allowed to build a unified model for tagging the elements of the WHO, WHAT and WHY dimensions of a US; each tag representing a concept with an inherent nature and defined granularity. Once tagged, the US elements can be graphically represented with an icon and the modeler can define the inter-dependencies between the elements to build one or more so-called Rationale Trees (RT). [Question/Problem] RT and their benefits have been illustrated on case studies but the ability to easily build a RT in a genuine case for software modelers not familiar with the concepts needs to be evaluated. [Principal ideas/results] This paper presents the result of a double exercise aimed to evaluate how well novice and experienced modelers were able to build a RT out of an existing US set. The experiment explicitly forces the test subjects to attribute a concept to US elements and to link these together. [Contribution] On the basis of the conducted experiment, we highlight the encountered difficulties that the lambda modeler faces when building a RT with basic support. Overall, the test subjects have produced models of satisfying quality. Also, we highlight these necessary conditions that need to be provided to the lambda modeler to build a consistent RT.

Keywords: User Story · Rationale Tree · Modeling experiment
Granularity

1 Introduction

In agile methods, requirements are often written through *User Stories (US)*. *User stories are short, simple descriptions of a feature told from the perspective of the person who desires the new capability, usually a user or customer of the*

system. US are generally presented in a flat list which makes the nature of the elements constituting them as well as their hierarchy and interdependence(s) difficult to evaluate [3].

The general US pattern relates a WHO, a WHAT and possibly a WHY dimension but, in practice, different keywords are used to describe these dimensions (e.g. Mike Cohn's *As a <type of user>, I want <some goal> so that <some reason>* [3]). Moreover, in the literature, no semantics has ever been associated to these keywords. Thus, Wautelet et al. [10] conducted research to find the majority of templates used in practice, sort them and associate semantics to each keyword. The key idea behind of [10] is that, using a unified and consistent set of US templates, the tags associated to each element of the US set provide information about both its nature and granularity. Such information could be used for software analysis, e.g., structuring the problem and solution, identifying missing requirements, etc. [5]. Most of the concepts of [10] are related to the i* framework [12], and a visual *Goal-Oriented Requirements Engineering (GORE)* model, the Rationale Tree (*RT*), has been formalized for graphical representation of US sets in [11]. Alternatively, a graphical representation using the use-case model based on the same concepts is proposed in [8].

A consistent RT allows its reader to identify the hierarchy of elements and their interdependence(s). This also provides a global view of the system to be developed. A RT is constructed from a tagged US set which is based on the US unified model proposed in [10]; of course a real life US set is seldom fully consistent so that when building a RT, the US set is sorted, cleaned, updated, etc. Concretely, the modeler, supported by a *Computer-Aided Software Engineering (CASE)* tool, should associate the tags to each US element and, in a visual window, link these US elements through means-end or traditional decompositions. When doing this, the modeler makes an assumption on (i) the nature of the US element (functional or not, coarse-grained or fine-grained) and (ii) how the US element needs to be fulfilled (immediately or by fulfilling other US elements found in other US or added to the set to ensure consistency). In terms of transformation of elements to a software design, the benefits of using the RT to build an agent-oriented architecture in [9].

This paper presents the results of an experiment where novice (students) and experienced (researchers) modelers are required to build RT out of two different US sets (cases). It is part of further validation of the applicability of the previously evoked research. The experiment has been designed in order to answer two main research questions. The first and main one is to see if, *starting from a US set, a lambda modeler is able to easily build a consistent RT*. The second is a side one and concerns *what are the necessary conditions to provide a lambda modeler the ability to build a consistent RT*.

2 Related Work

The need to test different decomposition techniques US with different agile methods and kind of stakeholders has been identified in [6]. We nevertheless only consider US as structured in the evoked form, independently of the agile method and

evaluate the perspective of the modeler only. Trkman et al. [7] propose an approach for mapping US to process models in order to understand US dependencies. Their approach is oriented to building an operational sequence of activities which is a dynamic approach not targeted to multiple granularity level representation. We, however, aim to build a rationale analysis of US elements which is a static approach allowing to represent and identify at once multiple granularity levels. Finally, as identified by [2], the representation symbols in a visual notation have an impact on the modelers understanding. We by default used the symbols of i^* but this parameter should be further studied.

3 Research Method: Designing the Modeling Experiment

The experimentation uses two cases: (**Case 1**) the carpooling system and (**Case 2**) the Book Factory. Due to the lack of space, we only expose **Case 2** here. The description of the **Case 1** can be found on p. 1 in an Appendix document placed online¹.

The Book Factory is a small Belgian retailer specialized in selling books, CD's and DVD's. The management has decided to invest in an online shopping environment for their customers in order to increase the customer-friendliness of their services. Within this online shopping environment, a user should have the possibility to place their orders online. Before an order is complete, a client should fill his online cart with products.

Secondly, the client should have to pay the invoice using an online payment. In order to be able to execute the payment, the system should calculate the invoice amount. Furthermore, the online payments are processed via the Ogone payment platform in order to increase the safety and security of the payment.

The related US set is provided within Table 1. A concept of the unified model has been associated to each US element; the interested reader can refer to [10] for their full definition. The type RT solution is provided in Fig. 1. This example is also the second exercise given to test subjects for the experiment. Let us note that, in US 5, *I need to calculate the total amount of the order* has been modeled as a *Capability* because it is seen as atomic while all the other elements tagged as *Tasks* are seen as being further decomposable. The choice can nevertheless be seen as arbitrary and as long as all of the elements we just referred to are tagged as *Task* or *Capability* the tagging can be seen as valid. Section 5 further discusses the interpretation and use of *Tasks* and *Capabilities*.

Process for Building the Modeling Experiment. As a first step in the research process, the different exercises (i.e. US sets to be tagged and transformed into a RT) used for the modeling experiment have been designed. In order to compare the output of the experiment subjects a *type solution* was built by a junior and 2 senior researchers (all part of the authors of the paper). Also, a theoretical part – to explain subjects the theory about the RT – has been built;

¹ <https://goo.gl/8ZT5tD> (this document is refereed to several times in the rest of this paper).

Table 1. US set in Case 2 of the feasibility study.

US ID	Dimension	User Story	Descriptive Concept Type
US 1	WHO	<i>As an owner</i>	Role
	WHAT	<i>I want my clients to be able to place orders online</i>	Hard-goal
	WHY	<i>So that the customer-friendliness of our services increases</i>	Soft-goal
US 2	WHO	<i>As a client</i>	Role
	WHAT	<i>I have to complete an order</i>	Task
	WHY	<i>So that I can place it online</i>	Hard-goal
US 3	WHO	<i>As a client</i>	Role
	WHAT	<i>I need to fill my 'online cart' with products</i>	Task
US 4	WHO	<i>As a client</i>	Role
	WHAT	<i>I need to pay my invoice</i>	Task
	WHY	<i>So that I can complete an online order</i>	Hard-goal
US 5	WHO	<i>As system component</i>	Role
	WHAT	<i>I need to calculate the total amount of the order</i>	Capability
	WHY	<i>So that the invoice can be paid</i>	Hard-goal
US 6	WHO	<i>As system component</i>	Role
	WHAT	<i>I want to pay my order online</i>	Task
	WHY	<i>So that my invoice is paid</i>	Hard-goal
US 7	WHO	<i>As a system component</i>	Role
	WHAT	<i>I need to process payments on the Ogone-payment platform</i>	Task
	WHY	<i>So that the payment is secured</i>	Soft-goal

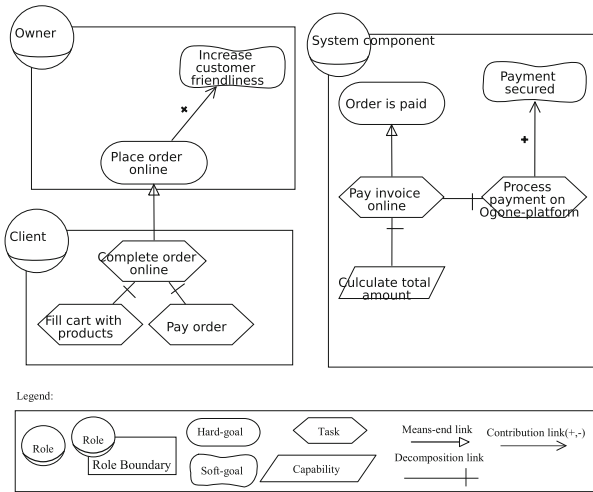


Fig. 1. Possible solution of Case 2 in the feasibility study.

it has been included in the set of papers given to test subjects. Finally, questions to measure some additional variables have been defined.

To evaluate the practical feasibility of the experiment, a primary evaluation/simulation with a group of researchers (PhD students and postdocs) at *Université catholique de Louvain (UCL)* has been done. Based on this test

feedback, some aspects in the layout of the modeling experiment have been changed/adapted. No content-related aspects have nevertheless been changed whereby the integrity of the evaluation basis between the first and second version of the modeling experiment has not been affected. Therefore, we also considered the data collected from this experimentation for analysis. The final version of the modeling experiment has been placed online².

Assignment and Measured Variables. Test subjects were asked to produce two separate US models based on two cases. These cases respectively consisted of a set of 4 and 7 US. The first US set was less complex than the second one in that the RT to build up was less complex. Since US and the production of a US-based model was new to the test subjects, the assignment has been split up in 5 steps, i.e.:

1. Identification of all elements within the WHO dimension of the US;
2. Identification of all elements within the WHAT and WHY dimension of the US;
3. Identification of the appropriate concept or tag (i.e., *Capability*, *Task*, *Hard-goal* or *Soft-goal*) for each element within the WHAT and WHY dimension of the US;
4. Graphical representation (and linking) of the US' WHAT and WHY elements;
5. Identification and representation of other links between the US elements.

Throughout the modeling experiment, additional questions have been asked in order to gather *additional variables* concerning the educational background, the tacit knowledge and the perception on difficulty of the different test subjects, i.e.:

- Their educational background (i.e., obtained diplomas);
- Their primary occupation (i.e., student, researcher, assistant, etc.);
- Their modeling knowledge (i.e., the modeling languages they already worked with);
- Whether or not they were familiar with GORE;
- Based on rating-scales, their knowledge on the i* framework and their knowledge concerning US as requirements artefacts within agile methods have been measured.

In between the different assignment steps (i.e., steps 1 to 5 as described above), the test subjects were asked to indicate their experience and perception concerning the understandability of the theory and concerning the difficulty of the steps to be executed. Latter elements have been measured using a rating-scale. At the end of the modeling experiment, some additional questions were asked in order to find out the global perceived experience of the test subjects when modeling the two cases. More specifically, they were asked to indicate which case was perceived as most difficult and, based on rating-scales, the global understandability of the proposed approach was measured.

² <https://goo.gl/i8GmJM>.

4 Data Collection and Participants' Modeling Knowledge

The experiments have been conducted with *three groups* of expertise. The first group consists of business students with a major in IT (known as *Business Students* in this paper). The second group consists of students in IT (known as *IT Students* in this paper). For the two former groups, the experiment has been done in class in the context of a special session of a compulsory course. The third group of expertise is made by the researchers of the pre-test (known as *Researchers* in this paper). For this last group, the experiment has been done in a single class room during working hours. The researchers participated on a voluntary basis; all of the researchers of the department were invited. These researchers all hold (at least) a master diploma with a major in IT. The use of three different groups of population notably allows us to *analyze the difference in execution of the assignment* and to study whether or not there are significant differences between these groups of various modeling experience. We nevertheless point out that all of the participants have chosen for a strong IT component in their present or past curriculum.

Since a concrete sample framework is lacking within the context of this modeling experiment, a non-stochastic sample method is used to compose the different samples. More precisely, the strategy of convenience samples has been used. Ultimately, three different samples have been composed. For the group of *Business Students*, the modeling experiment has been executed by 21 students within the master in Business Administration at KU Leuven campus Brussels. For the group of *IT Students*, the modeling experiment has been conducted with 35 students within the second bachelor Applied Informatics at Odisee campus Brussels. Finally, for the group of *Researchers*, the experiments have been conducted with 13 members of the academic staff of UCL.

The questions on background in Business Analysis shows that nearly the entirety of the participants have some preliminary knowledge in modeling. Indeed, only 2 out of 69 participants do not have such specific experience (i.e., 1 *IT Student* and 1 *Business Student*). They are able to model, at least one model, with the *Unified Modeling Language (UML)*, *Business Process Modeling Notation (BPMN)* or others modeling languages; but not GORE (only 2 *Researchers* have knowledge on GORE frameworks).

Concerning the question about knowledge of the i* framework by participants, results showed that none of them is an expert with the framework. Some students received a specific 2 h presentation on i* during another course. Nevertheless, over 50% never heard about it (but most are unaware that i* is GORE). Meanwhile, two thirds of the participants know what US are and some of them are experts in using them.

5 Tagging of User Story Elements

According to the US meta-model presented in [10], the US elements of WHO dimension can only be tagged as *Role*. No interpretation aspect need to be

Table 2. Tagging of the US elements in Case 1 and 2.

		Business Students					IT Students					Researchers				
		Task	Capability	Hard-goal	Soft-goal	Not present	Task	Capability	Hard-goal	Soft-goal	Not present	Task	Capability	Hard-goal	Soft-goal	Not present
Case1																
US2	WHAT	42.9%	33.3%	23.8%			31.4%	51.5%	11.4%	15.7%		53.8%	30.8%	15.4%		
	WHY			9.5%	90.5%				2.9%	97.1%			30.8%	7.7%	51.5%	
US3	WHAT	85.7%	14.3%				94.3%	5.7%				84.5%	15.4%			
	WHY	9.5%	4.8%	75.2%	9.5%		2.8%	8.5%	71.4%	14.3%	2.9%		76.9%	23.1%		
US4	WHAT	23.8%	76.2%				34.3%	62.8%	2.9%			30.8%	69.2%			
	WHY	38.1%	14.8%	47.6%	9.5%		48.6%	2.9%	37.1%	11.4%		46.2%	15.3%	15.4%	23.1%	
Case2																
US2	WHAT	85.7%	14.3%				66.7%	27.2%	6.1%			75.0%	8.3%	16.7%		
	WHY	4.8%	4.7%	81.0%	9.5%	9.0%	3.0%	66.7%	6.1%	15.2%		9.1%	90.9%			
US3	WHAT	52.4%	42.8%		4.8%		42.4%	36.4%	9.1%	12.1%		58.3%	25.0%	16.7%		
	WHY	4.7%		4.8%	90.5%					100%		8.4%	8.3%	83.3%		
US4	WHAT	66.7%	28.5%	4.6%			75.0%	18.8%	6.2%			91.7%	8.3%			
	WHY	52.4%	4.7%	42.9%			25.0%	18.8%	50.0%	6.2%		50.0%	16.7%	33.3%		
US5	WHAT	52.4%	47.6%				71.9%	25.0%	3.1%			41.7%	50.0%	6.3%		
	WHY	23.8%		76.2%			19.4%	29.0%	41.9%	6.5%	3.2%	27.3%	9.1%	36.4%	27.3%	
US6	WHAT	47.6%	42.9%	9.5%			57.6%	36.3%	6.1%			63.6%	27.3%	9.1%		
	WHY	14.2%	19.5%	66.7%	4.8%	4.8%	15.6%	12.5%	65.6%	6.3%		18.2%	9.1%	36.4%	36.4%	
US7	WHAT	47.6%	42.9%	9.5%			50.0%	37.5%	9.4%			45.5%	54.5%			
	WHY			100%					100%					100%		

legend: Highest occurrence within the sample in question

discussed here. US elements in the WHAT and WHY dimensions can nevertheless be tagged as *Capability*, *Task*, *Hard-goal* or *Soft-goal* and need to be discussed.

The results of US elements tagging for the WHAT and WHY dimensions of Case 1 and Case 2 are represented in Table 2. Since a valid interpretation for the first US of both cases was given as illustration they have been left out of the results.

Based on the information provided in Table 2, we can draw the conclusion that the tagging of the different US elements of both cases differs *within* as well as *between* the different samples. In other words, tagging of a US element as being *Capability*, *Task*, *Hard-goal* or *Soft-goal* cannot be characterized as being univocal (similar results have been highlighted by [1, 4] in the context of i* modeling). Also between the different samples, there are a lot of tagging discords; a few observations can be made:

- The tagging discords within the sample of *Business Students* is mainly between the tagging of a US element as being a *Task* or *Capability*. A higher variability in tagging of US elements can be observed within the samples of *IT Students* and *Researchers* especially within the Case 2;

- There exists some discords in what some US elements are. The confusion is mainly about the difference between a *Capability*, *Task* or *Hard-goal*;
- Despite this, test subjects unanimously agreed upon that the provided concepts (i.e., *Capability*, *Task*, *Hard-goal* and *Soft-goal*) were sufficient to model the different US sets. In other words, they did not witness having the need for additional concepts to accurately tag some US elements when performing the exercise.

As part of the modeling experiment, test subjects have been asked to indicate on a rating-scale whether or not the difference between the modeling concepts—i.e., respectively *Task* versus *Capability*, *Task* versus *Goal* (*Hard-goal* and *Soft-goal*) and *Goal* versus *Capability*—were clear. Within this rating-scale, the value 1 reflects the fact that the difference between the two modeling concepts was *not clear at all*. Conversely, the value 10 reflects a *complete awareness of the differences* between both modeling concepts. The descriptive statistics of these elements are provided in Table 3. Based on this data, we can draw the conclusion that, especially, the difference between *Task* and *Capability* was not completely clear for test subjects. Furthermore, the data indicates that *Task* and *Capability* were perceived as easier to differentiate from *Goal*.

Latter observation of the unclear difference between *Task* and *Capability* is confirmed by an analysis of the main modeling errors that have been made by the different test subjects. These modeling errors notably revealed that the atomic characteristic of *Capability* (i.e., the key feature that distinguishes *Capability* from *Task*) was not clear at all since a tremendous amount of test subjects graphically decomposed *Capability* elements into multiple sub-elements (using *decomposition-links*). This is not valid with respect to the presented base model. It, however, does not necessarily mean that the interpretation of element’s granularity is necessarily incorrect (this is evaluated through the quality of the RT in the next section).

Table 3. Understandability of the difference between the elements.

	Task vs. Capability			Task vs. Goal			Goal vs. Capability		
	Business Students	IT Students	Researchers	Business Students	IT Students	Researchers	Business Students	IT Students	Researchers
Average	5.52	6.12	5.54	7.14	7.21	6.23	7.29	6.91	6.46
Median	6	6	4	7	7	7	7	7	7
Minimum	2	1	1	1	2	2	2	2	1
Maximum	8	10	9	10	10	10	10	10	9

scale:

1 The difference between both elements is not clear at all

5 I'm not sure

10 The difference between both elements is completely clear

Next to this, a statistical test has been performed in order to test whether or not there exist significant differences within the samples in the test subject's 'understandability scores'. More specifically, the non-parametric Kruskal-Wallis test has been executed since the normality test (i.e., Kolmogorov-Smirnov) indicated that none of the variables involved were normally distributed. Latter non-parametric test verifies if multiple population variables have the same distribution. Based on the results of this test (not represented due to a lack of space, see the Appendix on p. 6) the conclusion can be drawn that no significant differences exist between the scores of *Business Students*, *IT Students* and *Researchers*³.

6 Analyzing the User Story Model with Rationale Tree

6.1 Global Evaluation of the User Story Model: Qualitative Approach

Business Students. The sample of students with an economical background succeeded rather well in producing a RT. However, the results showed that a few test subjects within this first sample tended at modeling each US separately instead of producing a global model for the complete US set in the cases. They failed in identifying corresponding elements within different US and they consequently modeled the same elements multiple times (i.e., one time per occurrence in a US). Latter observation nevertheless has to be put in some perspective in that it could possibly be correlated with one of the limitations of the modeling experiment. More precisely, since test subjects only received the minimal required amount of information for executing the assignment within the modeling experiment, one could argue that more information concerning the ultimate purpose of the graphical representation should have been depicted in more detail within the theory part of the modeling experiment. This probably could have resulted in a higher understanding of the primary rationale behind modeling US and could consequently have resulted in a higher ability to produce a RT of a US set. Another tendency that could be identified within the *Business Students* is that test subjects with a (basic) knowledge of US were able to make up a higher-quality hierarchical structure within their RT. Furthermore, analysis of the different models produced by the test subjects in all three samples revealed that, together with *IT Students*, *Business Students* tend to put a stronger emphasis on the process-related aspect of the US set in their model. Latter phenomenon could clearly be observed within the **Case 2**. For example, US3 and US4 respectively consist of the elements **Fill online cart** and **Pay invoice**. Both elements can be seen as sub-elements of the WHAT dimension in US2: **Complete an order**. Many students tried to model latter two elements (i.e., **fill cart** and **payment**) in such a way that the process-related sequence of these elements was represented in their model; i.e. that the result reflects the constraint that the online cart should be filled with products before the invoice

³ In the context of the statistical tests conducted in this work, a reliability of 95% has been used.

can be paid. Adjoining it, many test subjects within this first sample made the remark that some modeling elements were missing in order to represent **sequential conditions** between elements in the model.

IT Students. More than *Business Students*, *IT Students* failed in overviewing the ‘global model’ and tended to model each US separately. This resulted in the fact that their models consisted multiple ‘isolated’ elements without any link to another element. As a consequence, it is impossible to trace the dependency and hierarchy relationships between the different elements within the RT. One can thus state that *IT Students* were less able to produce a high-quality RT from a US set. A second observation that could be done is that the ‘technical’ background of the *IT Students* reveals itself within their different models. A few students namely modeled elements that were not part of the US set that has been included in the cases. These elements could commonly be categorized as more ‘technical’ elements that are part of the actual development of the systems. For example, some students represented an element **show ride** within their model of the **Case 1**. Others included the element **verify payment** within the boundaries of their model of the **Case 2**.

Researchers. Only taking into account the ability to produce a RT of a US set, one can state that *Researchers* produced higher-quality RT compared to students. In other words, *Researchers* were able to produce a better global model where the complete US set was represented in the RT. Within the models produced by the different test subjects in this sample, a tendency of modeling more elements than present in the US could be observed (i.e., elements that were not present in the US set). Furthermore, a lot of *Researchers* decomposed existing elements into (smaller) sub-elements. As an example, the WHAT dimension within US2 of the **Case 1** consists of the element **propose a ride from A to B with the price, location and time of departure, and number of seats available**. Instead of modeling this element as being one *Task*, many *Researchers* used 4 different elements to model this (i.e., one for **price**, one for **location**, one for **time of departure** and one for the **number of seats available**). Secondly, the different test subjects within this sample tended at identifying and modeling links that were outside the scope and boundaries of the definition that had been provided in the theory. More specifically, they used the broader definition of the links as present within the *i** framework.

Modeling Errors. Within the US models of the different test subjects, various modeling errors have been made. A frequently occurring modeling error concerned the decomposition of *Capabilities* into subcomponents. A second common error made by subjects of all three samples concerned the fact that the different roles (through boundaries) were not represented in the graphical US model.

Next to these modeling errors, nearly all US models of all test subjects contained one or multiple link errors (i.e., use of a faulty link). As an example, many test subjects in all samples used a *means-end link* between two *Tasks* while latter link has theoretically been defined as a link that is used between a *Task* and a *Hard-goal* if the former furnishes a realization scenario for the latter.

Table 4. Descriptive statistics of the number of elements and links modeled.

	Case1 Elements modeled			Case1 Links identified			Case2 Elements modeled			Case2 Links identified		
	Business Students	IT Students	Researchers	Business Students	IT Students	Researchers	Business Students	IT Students	Researchers	Business Students	IT Students	Researchers
Average	6.1	6.1	7.7	4.9	4.6	5.7	10.1	10.5	9.2	7.9	7.9	8.2
Median	6	6	6.5	5	4	4.5	10	9.5	9.5	8	8	9
Minimum	4	5	5	3	3	3	7	4	4	3	4	4
Maximum	11	9	17	10	8	13	13	13	13	11	13	10

The tremendous amount of linkage errors allows to draw the conclusion that some theoretical aspects concerning the different links have not been understood completely. This conclusion can directly be associated with the limited amount of information that has been given to test subjects.

Quantitative Evaluation of the US Models. Table 4 contains the data of the quantitative analysis of RTs. It allows to make a comparison between the US models made by the test subjects in the three different samples. Based on the results of the Kruskal-Wallis test (not represented here due to a lack of space, see the Appendix on p. 7), one can conclude that there are no significant differences between the number of elements and links modeled by the different test subjects in the three different samples. Latter non-parametric test has been executed since the Kolmogorov-Smirnov test has indicated that none of the variables involved are normally distributed.

6.2 Quoting the Performance in Modeling User Stories: Quantitative Approach

In order to be able to evaluate the individual performance of the test subjects in modeling the US sets in both cases, a score has been allocated to each US model. This score is notably based on three different evaluation criteria: *completeness*, *conformity* and *accuracy*.

Completeness has been used to verify whether or not all elements present in the different dimensions of the US set have been represented within the US model. For each element in the WHAT and WHY dimensions of a US that has been represented in the US model, 1 point was given.

In combination with completeness, the models have been evaluated with respect to conformity. During the exercises of the modeling experiment, the test subjects were asked to identify all elements in the WHAT and WHY dimension of the different US and classify each element as a *Task*, *Capability*, *Hard-goal* or *Soft-goal* (i.e., respectively steps 2 and 3 in the modeling experiment). In order

to verify if the appropriate modeling concepts have been used in accordance with the classification of the elements, the evaluation criterion of conformity has been used. More precisely, if there was conformity between the classification of an element and the modeling concept that has been used to represent that element, 0.5 points (per element) were given.

Based on the type solution of both cases, the fundamental links that should be present in the US models have been identified. More precisely, 4 fundamental links have been identified in the **Case 1** and 8 links in the **Case 2**. If one of the fundamental links was present in the US model of the test subjects, 4 points were given. If the link between the elements had been identified but the wrong type of link was used, only 1 point was given. This quotation of the ability of subjects to identify the links between the elements is the accuracy criterion.

Next to the scores on each evaluation criterion, a score on the global quality of the US models has been given. More specifically, an additional score on 10 was given for the **Case 1** and a score on 20 for the **Case 2**. The score on the global quality has been based on a general comparison of the US models with the type solution. Furthermore, additional factors have influenced the individual score of the global quality. More specifically, the fact that all *Roles* were correctly represented, the number of modeling errors and the quality of the RT were factors that have been taken into consideration in allocating the score on global quality. An overview of the different evaluation criteria and the allocated scores are provided in Table 5. Ultimately, a total mark on each case has been calculated based on the scores of the individual evaluation criteria. More precisely, a total score on 38 was given for the **Case 1** and a score on 73 was given on the second one. Both scores have eventually been reduced to a score on 10.

In order to get an overview of the ‘general performance’ of the test subjects in modeling the different US, a global score on 10 has been calculated. This score was based on the individual scores for **Case 1** and **Case 2**. Within the calculation of latter global score a weight of 30% has been allocated to the **Case 1** and a weight of 70% to the **Case 2**. The allocation of a different weight to both cases has been done since one could argue that a kind of ‘learning-effect’

Table 5. Evaluation criteria in quoting the US models.

Evaluation criterion	Allocated scores	Maximum score	
		Case 1 (4 US)	Case 2 (7 US)
Completeness	1 point per modeled element	8 points	14 points
Consistency	0.5 points per consistently modeled element	4 points	7 points
Accuracy	4 points per correct link (only 1 point if the wrong type of link is used)	16 points	32 points
Global quality	–	10 points	20 points

Table 6. General performance of modelers.

(a) Descriptive statistics of the global score.

	Business Students	IT Students	Researchers
Average	6.20	5.50	6.60
Median	6.60	5.30	6.50
Minimum	2.90	3.60	4.40
Maximum	8.30	7.40	8.60

(b) Averages Scores on Case 1 and 2.

Sample Groupe	Case 1	Case 2
Business Students	6.30	6.20
IT Students	5.60	5.40
Researchers	7.20	6.30

could have occurred after the execution of the **Case 1**. The **Case 2** furthermore consisted of a higher number of US, what implies that a bigger RT.

Table 6a consists of the descriptive statistics of the global score (on 10) that measures the performance of the test subjects in modeling a set of related US. The normal distribution of this global performance score⁴ allows to perform the ANOVA-test in order to verify if there exists some significant differences between the scores of the different samples (i.e., *Business Students*, *IT Students* and *Researchers*). Based on the results of this test (not represented here due to a lack of space, see the Appendix on p. 8), the conclusion can be drawn that there indeed exist significant differences between the scores of the different test subjects in the three samples. More precisely, the results of the post-hoc test of Bonferroni (not represented here due to a lack of space, see the Appendix on p. 9) learn that, with a reliability of 95%, a significant difference can be found between the scores of the *IT Students* and those of the *Researchers*. There is no significant difference between the scores of *Business Students* and *IT Students* and between those of *Business Students* and *Researchers*.

Next to the differences in the global score between the three samples, one could question whether there exists a significant difference in the individual performance of modeling both cases. Table 6b represents the average score on both cases per sample. In order to test for significant differences in the score of **Case 1** compared to the score **Case 2**, the paired samples t-test is performed on the different scores of each particular sample. The results of these tests (not represented here, see the Appendix on p. 10) show that no significant differences can be identified in the performance of *Business Student* and *IT Students* in modeling the US sets in both cases. This contrary to the sample of *Researchers*, where can be concluded (with a reliability of 95%), that the scores on **Case 1** significantly differ from those of **Case 2**.

7 Analyzing the Experience of Test Subjects

7.1 Evaluating the Understandability of the Theory

In order to measure the understandability of the theory, four questions have been asked to test subjects. These questions were to be answered using a rating-scale

⁴ Both the Kolmogorov-Smirnov test as well as the Shapiro-Wilk test indicated that the variable of the global score was normally distributed.

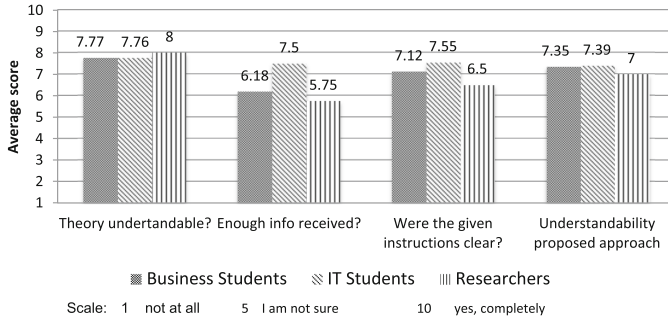


Fig. 2. Understandability of the theory.

going from 1 for *not at all* to 10 for *completely*. A first question concerned the understandability of the introductory theory part of the modeling experiment. Secondly, test subjects were asked if they received enough information to produce the models. Thirdly, they were also asked if the given instructions to model the US sets were clear. The fourth question concerned the understandability of the proposed approach for producing a US model using a RT. The average score of these questions are represented within Fig. 2.

Analysis of the results of these additional questions reveals that, despite the fuzzy differentiation between *Task* and *Capability*, the theory was rather understandable for most test subjects. However, an evaluation of the most common modeling error shows that not all aspects within the theory have been understood completely. In all three samples, a considerable amount of test subjects made particular modeling errors from which latter conclusion can be derived. As stated within Subsect. 6.1, a tremendous amount of modeling errors concerned the fact that *Capabilities* have been graphically decomposed into multiple sub-elements. This shows that the atomic characteristic of a *Capability* (i.e., the key feature that distinguishes it from a *Task*) has not always been understood. Another common modeling error – several elements were linked to a *Hard-goal* by means of a *means-end link* – allows to draw the conclusion that the theoretical definition of this type of link has not always been understood properly.

7.2 Evaluation of the Perceived Difficulty

A last component within the analysis of the results the modeling experiment concerns an evaluation of the perceived difficulty by the different test subjects in the three samples. Within the modeling experiment, several variables have been included in order to be able to measure the perception of the test subjects on the difficulty. The perceived difficulty has in fact been measured on three different levels. On a first level, the test subjects were asked to indicate on a rating-scale their perceived degree of difficulty in modeling the two cases. Secondly, the test subjects have been asked for their experience in executing the different steps (i.e., steps 1 to 5, see Sect. 3). On a third level, they were asked to indicate if **Case 1**

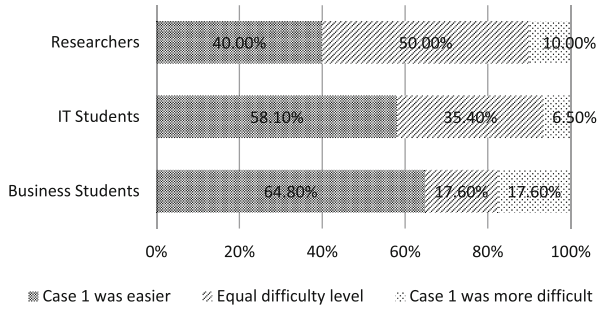


Fig. 3. Graph difficulty Case 1 versus Case 2.

was easier, of an equal difficulty level or more difficult to model compared to Case 2 (as can be seen in Fig. 3).

Perceived Difficulty to Model both Cases. The first variable that has been used to measure the perceived difficulty concerned the perception on the global difficulty to model the US sets in Case 1 and 2. More precisely, the test subjects were asked to answer the question ‘*was it difficult to model both cases?*’ on a rating-scale. On this scale, the value 1 represented the answer *not at all* and the value 10 represented the answer *yes, completely*. The average score given by the different test subjects on this question is 4.76 for Business Students, 5.52 for IT Students and 5.75 for Researchers. In order to be able to provide an answer to the question if there exist some significant differences between the perceived difficulty by *Researchers*, *IT Students* and *Business Students*, the non-parametric Kruskal-Wallis test has been performed. The results of this test (not represented here due to a lack of space, see the Appendix on p. 12) indicate that there exists no significant difference between the global perceived difficulty to model the US in both cases.

8 Lessons Learned and (CASE Tool) Enhanced Support

As a second research objective, we want to identify the necessary conditions that need to be provided to the lambda modeler to build a consistent RT. We have, indeed, modified the CASE tool (see [9] for an explanation of the different available views, all views are always kept consistent) that has been built to support the creation of RT in order to better support the RT modeling activities. This way, we aim to help the modeler in the modeling process and avoid him to make some mistakes.

RT Validity. In order to deal with the ambiguity between the *Task* and *Capability* elements, we have included a model checker functionality that, when loaded (clicked upon), evaluates if all the leaf nodes of the RT and only these are tagged as *Capabilities*. If it is not the case, the modeler receives a warning together with some theoretical explanations and is invited to modify the associated tag. If the

element's tag is modified, the icon is updated. He nevertheless still have the ability to refuse the change (it can indeed be that the element is a Task but its decomposition has not been done yet).

Completeness Aspect. One of the aims of a RT is to be able to study the completeness of requirements depicted in the US set through decomposition. As seen, the modeler as a natural tendency to try to add missing elements to complete the requirements model. Missing elements can be easily added in the RT. When using the model checker the modeler is explicitly shown dependencies with one leaf to invite him completing missing elements. Also a process view has been included allowing to model elements in a sequential order using BPMN; *Task* elements can be included as BPMN *sub-processes* and *Capability* elements as *activities*. This however remains an option.

Constraint Checking. Finally, to deal with the difficulty some modelers may have to link elements, we are developing an algorithm that automatically builds clusters of US elements in function of their semantic relatedness. Then, the modeler can make use of these clusters to link elements. The effectiveness of this method nevertheless still needs to be studied/validated.

9 Threats to Validity, Limitations and Conclusion

The first and main threat to validity comes from the quoting system itself. The latter has been built through an analysis of type solutions with the aim to define the criteria making these models of high quality. While we have justified the importance of the used criteria for the overall model evaluation, others could have been included but, more importantly, their balance – determined by the involved researchers themselves – can be seen as arbitrary. This issue could be further investigated in two ways. Firstly, we can make an independent study to (re)determine the evaluation criteria and their balance by, for example, asking the opinion of agile experts and practitioners. Concretely we can submit to experts RT built out of sets of US from cases they are familiar with. Then, we can ask them about their quality to determine what criteria they consider/use for evaluation and the relative importance of these criteria. With this new evaluation framework we can then reexamine the scores and results. A second way can be to use the criteria we have already used but to make variations in their relative weights to see how it impacts the overall scores and results.

A second threat to validity comes from the relatively small size of the US samples; respectively 4 and 8 US. One could not immediately generalize the results to a case with a significantly higher number of US. Another modeling experiment will be conducted to evaluate the capacity to build a RT out of a large US sample (over 20 US). Variants in the experiment will also include missing elements in requirements from the US set to evaluate if the RT helps with their identification.

The first limitation concerns the fact that the different test subjects only received a limited amount of information concerning the proposed approach of

modeling US. To keep the time required to complete the modeling experiment within acceptable boundaries, only the minimal required information on modeling constructs (i.e., *Task*, *Capability*, *Hard-goal* and *Soft-goal*) and the different links between these elements have been included within the theory section of the experiment. In an ideal situation, more information/details on US and on the graphical notation should have been given.

The second limitation concerns the size of the different samples. There has been a large difference between the number of test subjects within each individual sample. Furthermore, the size of the samples (especially the sample of *Researchers*) is rather small what limits the ability to reflect the results from the study towards the scope of the complete population with an acceptable reliability level. The lack of professionals very familiar with US as test subjects is a third limitation that can be identified.

An evaluation of the interpretation of the different elements in the WHAT and WHY dimension of a US set has shown that there existed some discord in the classification of the elements. Two possible reasons for latter discord can be identified. Firstly, particular elements allow by nature to be interpreted in several ways. On a second level, the interpretation discords are a direct consequence of the lack in understanding the theoretical differences between the various elements. This is primarily the case for a *Task* and a *Capability*. These conclusions are confirmed by analyzing the most common modeling errors, where a tremendous amount of test subjects graphically decomposed a *Capability* into multiple sub-elements. Despite the interpretation differences in the modeling experiment, the large majority of test subjects agreed upon the fact that no additional concepts (next to the ones of a *Task*, a *Capability*, a *Hard-goal* and a *Soft-goal*) are required to represent the US elements.

Concerning the ability to build up a RT, most of the test subjects were able to produce an acceptable model out of a US set. The different students however tended at modeling each US separately. This notably resulted in a model with multiple ‘isolated’ elements that have not been linked to other ones. Students furthermore have put a stronger emphasis on the process related sequence of the elements. Some of them argued that the model should contain specific modeling elements to represent process-related sequence of the different elements. *Researchers* by contrast tended at modeling additional elements that were not represented within the set US.

Even if the assignment of modeling two US sets has been perceived as quite difficult by the different test subjects, we showed that with minimal or no knowledge of GORE, people have been able to build a visual representation of a US set through a RT with minimal theoretical explanations. The identification of the different links has been perceived by the test subjects in all three samples as being the most difficult; this is nevertheless more related to domain knowledge and further analysis than to the transformation of the US set in the RT as such. The application of the method on a large US set in a professional IT context has since then also been realized. The application/interpretation of theory has there not been reported as an issue and multiple new benefits of the RT in an agile context have been identified.

References

1. Abad, K., Pérez, W., Carvallo, J.P., Franch, X.: *i** in practice: identifying frequent problems in its application. In: Proceedings of the 32nd ACM Symposium on Applied Computing (2017)
2. Caire, P., Genon, N., Heymans, P., Moody, D.L.: Visual notation design 2.0: towards user comprehensible requirements engineering notations. In: 21st IEEE International RE Conference, Rio de Janeiro-RJ, Brazil, pp. 115–124. IEEE Computer Society (2013)
3. Cohn, M.: *Succeeding with Agile: Software Development Using Scrum*, 1st edn. Addison-Wesley Professional, Boston (2009)
4. Dalpiaz, F.: Teaching goal modeling in undergraduate education. In: Proceedings of the 1st International iStar Teaching Workshop, CEUR Workshop Proceedings, vol. 1370, pp. 1–6 (2015)
5. Liskin, O., Pham, R., Kiesling, S., Schneider, K.: Why we need a granularity concept for user stories. In: Cantone, G., Marchesi, M. (eds.) XP 2014. LNBIP, vol. 179, pp. 110–125. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-06862-6_8
6. Taibi, D., Lenarduzzi, V., Janes, A., Liukkunen, K., Ahmad, M.O.: Comparing requirements decomposition within the scrum, scrum with kanban, XP, and banana development processes. In: Baumeister, H., Lichter, H., Riebisch, M. (eds.) XP 2017. LNBIP, vol. 283, pp. 68–83. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57633-6_5
7. Trkman, M., Mendling, J., Krisper, M.: Using business process models to better understand the dependencies among user stories. *Inf. Softw. Technol.* **71**, 58–76 (2016)
8. Wautelet, Y., Heng, S., Hintea, D., Kolp, M., Poelmans, S.: Bridging user story sets with the use case model. In: Link, S., Trujillo, J.C. (eds.) ER 2016. LNCS, vol. 9975, pp. 127–138. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47717-6_11
9. Wautelet, Y., Heng, S., Kiv, S., Kolp, M.: User-story driven development of multi-agent systems: a process fragment for agile methods. *Comput. Lang. Syst. Struct.* **50**, 159–176 (2017)
10. Wautelet, Y., Heng, S., Kolp, M., Mirbel, I.: Unifying and extending user story models. In: Jarke, M., Mylopoulos, J., Quix, C., Rolland, C., Manolopoulos, Y., Mouratidis, H., Horkoff, J. (eds.) CAiSE 2014. LNCS, vol. 8484, pp. 211–225. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07881-6_15
11. Wautelet, Y., Heng, S., Kolp, M., Mirbel, I., Poelmans, S.: Building a rationale diagram for evaluating user story sets. In: 10th IEEE International Conference on Research Challenges in Information Science, RCIS 2016, Grenoble, France, 1–3 June 2016, pp. 477–488 (2016)
12. Yu, E., Giorgini, P., Maiden, N., Mylopoulos, J.: *Social Modeling for Requirements Engineering*. MIT Press, Cambridge (2011)