Evaluating the Impact of User Stories Quality on the Ability to Understand and Structure Requirements

Yves Wautelet, Dries Gielis, Stephan Poelmans, Samedi Heng

Evaluating the Impact of User Stories Quality on the Ability to Understand and Structure Requirements

Yves Wautelet[1], Dries Gielis[1], Stephan Poelmans[1], and Samedi Heng[2]

[1] KU Leuven, Leuven, Belgium
{yves.wautelet, stephan.poelmans}@kuleuven.be,
[2] HEC Liège, Université de Liège, Liège, Belgium
samedi.heng@uliege.be

Abstract. Scrum is driven by user stories (US). The development team indeed uses, to fill the project's and the sprints' backlog, sentences describing the user expectations with respect to the software. US are often written "on the fly" in structured natural language so their quality and the set's consistency are not ensured. The Quality User Story (QUS) framework intends to evaluate and improve the quality of a given US set. Other independent research has built 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 granularity. Once tagged, the US elements can be graphically represented through an icon and the modeler can link them when inter-dependencies are identified to build one or more Rationale Trees (RT). This paper presents the result of an experiment conducted with novice modelers aimed to evaluate how well they are able to build a RT out of (i) a raw real-life US set (group 1) and (ii) a new version of the US set improved in quality using QUS (group 2). The experiment requires test subjects to identify the nature of US elements and to graphically represent and link them. The QUS-compliant US set improved the ability of the test subjects to make this identification and linking. We cannot conclude that the use of the QUS framework improved the understanding of the problem/solution domain but when a QUS-compliant US set is used to build a RT, it increases the ability of modelers to identify Epic US. Building a RT thus has a positive impact on identifying the structure of a US set's functional elements.

Keywords: User Stories; Rationale Tree; Quality User Story; Modeling Experiment

1 Introduction

Agile methods often describe software requirements with *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 and structure of the elements constituting them difficult to evaluate [3]. Commonly, US templates relates a WHO, a WHAT and possibly a WHY dimension and 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> [2]). In the literature no semantics have been associated to these keywords. Thus, Wautelet et al. [9] collected the majority of templates used in practice, sorted them and associated semantics to each keyword. The key idea is that, using a unified and consistent set of US templates, the tags associated to each element of the US set provide information about its nature and granularity. Such information could be used for software analysis, e.g., structuring the problem and solution, identifying missing requirements, etc. Most of the concepts of [9] are related to the i* framework [12] so that a visual *Goal-Oriented Requirements Engineering* (*GORE*) model, the *Rationale Tree* (*RT*), has been formalized for graphical representation of US sets in [8,10].

In parallel, Lucassen et al. [4] have proposed the Quality User Story (QUS) framework, a linguistic approach to evaluate and improve the quality of individual US and US sets. US are often written with poor attention and their quality can be improved by applying a set of 13 criteria. QUS is supported by the *Automatic Quality User Story Artisan* (*AQUSA*) software tool. Based on natural language processing techniques, AQUSA detects quality defects and suggests remedies. Domain experts also need to be involved in the US quality improvement process to fine tune the US set. Overall, a QUS-compliant US set is aimed to enhance readability and better support the human understanding of the software problem and solution than its non-compliant counterpart; this further helps stakeholders during all of the software development activities.

Even if they are basically independent researches, an experiment has been conducted to test whether the usage of the QUS framework leads to a US set allowing a modeler to build a RT of higher quality than one that would have been built with the original US set. For this purpose, a real-life US set has been selected and enhanced in quality using the QUS approach with the help of the AQUSA tool and domain experts (we have then a "raw" and a QUS-compliant US set). Students from the master in Business Administration (with a major in IT and familiar with various modeling techniques) at KU Leuven campus Brussels have served as test subjects. A first group was required to perform small exercises and build a RT out of the raw US set, the second one out of the QUS-compliant US set. The difference in quality of the RTs built and their constituting elements' relevance are studied in this paper.

2 Related Work

The need to test different decomposition techniques of US with different agile methods and kinds of stakeholders has been identified in [6]. In this paper we only consider US as structured in the Cohn's form, independently of a specific 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 levels representation. We, however, aim to build a rationale analysis of US elements which allows to represent and identify at once multiple granularity levels but does not show explicitly the sequence of activities. As identified by Caire & al. [1], 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 could be further studied.

Wautelet et al. [11] made an experiment using the unified model of [9] 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 were graphically represented by building one or more RTs. The research consisted 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 forced the test subjects to attribute a concept to US elements and to link these together. On the basis of the conducted experiment, difficulties that the modeler faces when building a RT with basic support were identified but overall the test subjects produced models of satisfying quality. The experiment of Wautelet et al. [11] can be seen as preliminary to the one conducted in this paper. We indeed here also guide subjects into the tagging of US elements and build a RT out of US sets. The main innovation here is that there is a variation of quality among the US sets submitted to the subjects.

3 Research Approach and Background

3.1 Research Hypothesis and Goals

Research Hypotheses According to Lucassen et al. [5] the use of the QUS framework effectively decreases the quality defects within US. One of the main expectations towards the use of the QUS framework in the experiment is thus that the quality (evaluated by scores) of the represented RTs will be higher with the QUS-compliant US set. This specifically means that we expect an improvement in identifying relevant **software functions** and elements like **Epics, Themes, Non Functional Requirements (NFRs)** and possibly **missing requirements**. The interference of the RT to identify the concepts is expected to be positive, especially for Themes and Epics because their identification is specifically supported by the RT. The **goals** of the experiment are then:

- To analyze the ability of the subjects to understand and identify different concepts (NFRs, missing requirements, Epics & Themes) related to US sets;
- To analyze and verify the ability of the subjects to build a RT from a set of US taken from a real-life case;
- To analyze the impact of the RT on the subjects' ability to identify and distinguish the previously mentioned concepts related to US sets;
- To analyze and measure the impact of the QUS-compliant US set on (i) the ability of the subjects to identify and distinguish the nature and granularity of elements present in US (before and after the use of the RT) and (ii) to build a RT.

3.2 Building the Experiment

A BPMN workflow of the followed research steps can be found in Appendix C[3]. We have created two versions of the experiment and randomly divided the subjects in two groups. One group that receives the experiment with the "raw" US set (available in Appendix A & B), and the other group that receives the experiment with the "QUS-compliant" set (also available in Appendix A & B).

[3] All appendices are available at: http://dx.doi.org/10.17632/st8byw8hkz.1

The real-life US set has been furnished by an organization that wants to remain anonymous; it is called "Company X" here. The latter furnished a document with US sets concerning the development of a web-application. From the original document, 2 (raw at this stage) US sets were selected (1 set for each part of the experiment). Then, several exercises were built together with theoretical explanations and instructions.

Fabiano Dalpiaz, involved as a promoter in the development of the QUS framework, ran the raw US sets through the AQUSA tool and delivered the generated reports. The tool does not include all the criteria so that a manual tagging was done by Fabiano to evaluate the US sets based on all the criteria (see Appendix D). Fabiano also added some comments to some of his tags to clarify his answer (Appendix D). Note that tagging a US means here to answer "yes" or "no" to the 13 criteria. The research team then met with an IT manager and a developer of company X to re-discuss and improve the US set. Both employees were involved in writing the US; they clarified some aspects allowing to build the final version of the two QUS-compliant US sets. With the raw and QUS-compliant US sets at disposal, the final version of the experiment was discussed by the research team. Based on this, some layout was changed and more context and explanations were added to the experiment document.

The last step was to create a well-founded solution. Each of the research team members created individually a possible solution for the RT. These solutions were compared among each other and discussed. After that, a joint solution that became the "moving golden standard" was set-up, meaning the solution of the RT could evolve during the corrections of the experiment. Indeed, when a subject modeled an element or link that was valid but not considered previously, it could be added to the solution after discussion among the research team members. The solutions of the exercises of both groups, with the moving golden standard of the RT included, are shown in appendix G & I. Appendix E contains a timetable that gives an overview of the iterations that were made to conduct the experiment. Each time information about when and how the meeting took place, who was involved and what the outcome was, is given. The conducted experiments of both groups are depicted in Appendix F (group 1) & H (group 2), followed by the solutions of the experiments in Appendix G (group 1) & I (group 2).

3.3 Assignment and measured variables

In an introductory part, questions have been asked to gather additional information that could be used as variables for the analysis. The following list of questions were asked: (i) educational background; (ii) primary occupation (student, researcher, teacher, ...); (iii) experience with software modeling (Likert-scale from 1 to 5): if they had experience, we further asked what languages they worked with; (iv) amount of years of experience with software development; and (v) 8 Likert-scale questions, from 1-5, about their knowledge of US, User Story Mapping (USM), NFR, US as requirements in agile methods, Epics, Themes, missing requirements and Entity Relationship Diagram (ERD).

Exercises part 1: The exercises for the first part consist in the identification of the following concepts: (i) non-functional requirements (exercise 1); (ii) Epics (exercise 2); (iii) Themes (exercise 2 (as well)); and (iv) Missing requirements (exercise 3). The subjects received some context information about the application to develop together

with a reference in the document's appendix where a list of US-related concepts were explained. The exercises of part 1 were based on the first US set of Company X. The first US set consists in 13 US in its "raw" form (thus for group 1, see Appendix A) and in 11 US in its QUS-compliant form (thus for group 2, see Appendix A). The entire sets were nevertheless split into small samples for the needs of each exercise containing 3 to 4 US. After having made the exercises, the subjects were asked to quantify, by using a Likert-scale from 1 to 5, the clarity of the explanations of the concepts and the difficulty they perceived in identifying these concepts.

Exercises part 2: The exercise for the second part consists in one global modeling exercise to build a RT. Theoretical background about the different types of elements (i.e., *role*, *task*, *capability*, *hard-goal* and *soft-goal*) and links (i.e., *means-end*, *decomposition* and *contribution*) used in the RT was given together with a running example of 4 US. The exercise of part 2 is based on the second US set of Company X. The latter US set consists in 7 US in its "raw" form (thus for group 1, see Appendix B) and in 7 US in its QUS-compliant form (thus for group 2, see Appendix B).

The subjects received information about the context of the application development in company X together with a second set of US. Based on a study by Wautelet et al. [11], the test subjects had to execute the following steps to model a RT (see the experiment document in the Appendix F for group 1 and Appendix H for group 2):

- **Step 1**: Identify the WHO element from each US;
- **Step 2**: Identify the elements from the WHAT- and WHY-dimension in every US;
- **Step 3**: Identify, for each element of the WHAT- and WHY-dimension, the construct that will be used for their graphical representation, according to the theory;
- **Step 4**: Graphically represent all elements identified in steps 2 & 3 and create a RT by linking them;
- **Step 5**: Identify the possible missing links to complete the graphical representation.

For steps 2 and 3, the first US was given as an example. The subjects were asked to identify the same concepts as in part 1 but this time using the RT to support them in the process. Note that the ability of identifying a NFR was not explicitly asked again because it was implicitly included in the modeling exercise. The last part consisted in 4 Likert-scale questions about the understandability and easiness of using the RT.

3.4 Data collection

To collect the data, the experiment has been executed by 34 Business Administration students with a specialization in Business Information Management at the KU Leuven campus Brussels. Before the start of the experiment, Yves Wautelet gave a 30 minutes introduction about US to both groups at the same time. The subjects were then divided in two groups of 17, one that used the raw set as input and the other that used the QUS-compliant one. The subjects were randomly divided by "blindly" giving them a piece of paper on which "1" or "2" was written. Subjects that received "1" stayed in the same room and were given the experiment with the raw set. Subjects that received "2" had to go to a second room where they were given the experiment with the QUS-compliant set.

3.5 Evaluating the Experiment's Results

The solutions used to evaluate the subject's representations are depicted in Appendix G for group 1 and Appendix I for group 2. Due to their small size, the solutions for the exercises in part 1 did not lead to much discussions and were rapidly adopted. For the large exercise of part 2, the solution is based on a "moving golden standard". Although all of the solutions are highlighted within the appendices, some more explanation about the RT of part 2 should be given. The research team chose to distinguish three hard-goals within the solution that were all separately connected with a task by a means-end link. The following three hard-goals were chosen because they all express a coarse-grained functionality: *Correct errors in personal information*, *Sign in with user account*, and *Register myself*. Besides identifying those elements, the test subjects also had to identify Epics, Themes and missing requirements using of their RT. *As an End User I want to register myself So that I can sign in with a user account* is considered an Epic US. The US indeed contains clear high-level elements while US 2, US 4 and US 5 are related to US 3.

4 Analyzing the results of the experiment

4.1 Preparing the data for analysis

Data was analyzed with SPSS. Variables have been defined and it has been ensured that their results could be compared by rescaling their total score. The latter was done because there was a difference in the value of the total score within some exercises between the experiment in group 1 and 2. Also, the relevant variables have been put in percentages so scores from different exercises within and between groups could be compared in a consistent way. The next step has been to evaluate and define useful factors. A short description of the used variables is given hereafter.

Description of the variables As previously mentioned, an introductory part of the experiment document given to the subjects collected some additional information about them. The variables that were collected are the following:

- EduBackground: highest education level obtained (high school, bachelor, master);
- Experience: the experience in software modeling (Likert-scale[4]);
- KnownModelingLanguage: what modeling languages they have experience with;
- MonthsOfExperience: how many months of experience with software development, regarding any method or technique;
- KnowledgeUserStories: their knowledge about US (Likert-scale);
- KnowledgeUserStoryMapping: their knowledge about USM (Likert-scale);
- KnowledgeNFR: their knowledge about NFRs (Likert-scale);
- KnowledgeUSInAgile: knowledge about US as requirement artifacts in agile software development methodologies (Likert-scale);

[4] A Likert-scale from 1-5 that goes from "never heard of it" to "expert in topic", is used in every variable with a "Likert-scale".

- KnowledgeEpicUS: their knowledge about Epic US (Likert-scale);
- KnowledgeUSThemes: their knowledge about Themes in US (Likert-scale);
- KnowledgeMissingRequirements: their knowledge about MR (Likert-scale);
- KnowledgeERD: their knowledge about Entity-Relationship (Likert-scale).

The variables that measure the score of the subjects on the different exercises were named *ScoreNFR*, *ScoreTheme*, *ScoreEpic* and *ScoreMR*. A distinction was made between the exercises of parts 1 and 2.

The ability of the subject to identify a NFR in part 2 was a part of the exercise on the RT and was named *ScoreSoft_Goal*. After the exercises, the subjects' perception on their ability to solve the exercises was asked and transformed into variables *DifficultyNFR*, *-Themes*, *-Epics* and *-MR* for part 1 as well as *FindMR*, *-Epic* and *-Theme* for part 2. The perception of the subjects' ability to identify soft-goals[5] in part 2 was not asked explicitly because it was captured in the perception of modeling the overall diagram. After the first part, the subjects were also asked to give their perception on the understandability of the concepts explained, respectively named *UnderstandUS*, *-NFR*, *-Epic* and *-Theme*. While Epics and Themes are related concepts, *ClearDifferenceEpic_Theme* asked whether the difference between both concepts was clear.

The modeling exercise in part 2, regarding the RT, measured the ability of the subjects to model each construct separately. *ScoreCoarseGrainedFunctionality*, *-Hard_Goal*, *-Soft_Goal*, *-Task*, *-Capability*, *-Links*, *-ConsistentTree* and *-MissingLink* were used as variables to measure their performance. Subjects received points on their ability to identify the coarse-grained functionalities from the US. They also received points when they indicated these functionalities as hard-goals, could identify the soft-goals, tasks and capabilities and connect the relevant elements by using the correct links. The RT was also analyzed on its consistency and could be divided into 3 levels. A consistently modeled RT was considered *a clear hierarchical structure were most of the relevant elements were linked*, subjects received the full points in this case. A partially modeled RT combines *at least 2 different US with no clear hierarchical structure*; this was given half of the points. A graphical model were *no US were linked*, was given 0.

After the exercise, a few questions about the use of the RT were asked. *HelpTree_MR*, *-_Epic*, *-_Themes* are the variables that captured the perception of the subject on how the RT helped in identifying the concepts. To end the experiment, 4 variables about the subjects' perception on the RT: (i) *IntroTree_Clear_Understandable* and *TheoryElementsLinks_Clear_Understandable*, measured how clear and understandable the introduction and the theory about the different elements and links was; (ii) *SkilfulAtUsingTree* measured whether the subjects would find it easy to become skilful at using the RT; and (iii) *ApplyTreeDailyWorkLife* measured whether the subjects would find it easy to apply the RT in their daily work life to evaluate US sets. The perceptions, mentioned above, were measured by a Likert-Scale from 1 to 5 where 1 means "Not at all" and 5 means "Extremely".

Factor Analysis A Principal Component Analysis was executed to reduce the amount of unstructured information from variables that are associated with a common latent

[5] Typically in the RT, a NFR is represented as a softgoal so that, in the rest of this paper, every time we refer to softgoal we implicitly mean a NFR.

(i.e., not directly measured) variable. Table 1 shows the relevant factors that were found and used during the analysis of the results. A total of six factors was found, Appendix J shows which items are related to which factors within the component matrix. The table shows all factors were usable because they all had an acceptable Kaiser-Meyer-Olkin (KMO) test (above 0,5). Besides that, the Bartletts Test of Sphericity was significant in every factor. Every factor had a sufficient percentage of total variance explained and a reliability analysis showed the Cronbach's alpha was high enough (above 0,6).

Table 1. Factor analysis.

Factors	**Factor loadings of items**	**KMO**	**Total variance**	**Cronbach's alpha**
F1: KnowledgeUS	**KU1**: 0,845; **KU2**: 0,785; **KU3**: 0,676; **KU4**: 0,672	0,722	55,987	0,731
F2: KnowledgeMacroConceptsUS	**KMC1**: 0,900; **KMC2**: 0,935; **KMC3**: 0,742	0,619	74,470	0,806
F3: UnderstandabilityMacroConcepts	**UMC1**: 0,718; **UMC2**: 0,731; **UMC3**: 0,920; **UMC4**: 0,920; **UMC5**: 0,607	0,796	62,246	0,797
F4: EasinessMacroConcepts_Part2	**EMC1**: 0,709; **EMC2**: 0,910; **EMC3**: 0,895	0,627	71,051	0,792
F5: HelpOfTreeMacroConcepts_Part2	**HTMC1**: 0,890; **HTMC2**: 0,870; **HTMC3**: 0,891	0,733	78,085	0,857
F6: ClearnessEasinessOfUseTree	**CET1**: 0,826; **CET2**: 0,768; **CET3**: 0,885; **CET4**: 0,844	0,751	69,211	0,848

4.2 A between-group comparison: analyzing the impact of the QUS framework

The first comparison that is made is the between-group one. The different scores on the exercises are compared by testing whether there is a significant difference between the means of group 1 and 2. In that way, there will be checked whether the use of a QUS-compliant US set improves the ability of the subjects to identify the different concepts before and after using the RT and improves the ability to build a graphical representation.

The **experience in software modeling** of respondents has also been analyzed. Due to a lack of space and because it is not fundamental for the overall understanding of the paper, it has been placed in Appendix K.

Analyzing the scores In this section some analysis regarding the scores of the exercises will be compared between both groups to check whether the QUS framework had a possible effect on the scores. Table 2 shows the overall scores of the exercises in part 1, part 2 and the modeling exercise of the RT.

The variables that are included in the overall scores are the following: *ScoreNFR* (*ScoreSoft_Goal* for part 2), *-Theme*, *-Epic* and *-MR*. The exercises concerning the latter

concepts can be found in Appendix F and H. The overall score of the modeling exercise is the sum of the scores of all separate elements that had to be modeled. As mentioned previously, the scores are expressed as percentages for consistency reasons.

Table 2. Comparing the means of the overall scores.

Variables	Mean group 1	Mean group 2	Mean difference (% points)
Percentage of score of the exercises in part 1	64,71	51,70	13,01*
Percentage of the score of the exercises in part 2	49,91	57,62	7,71
Percentage of the score of modelling the Rationale Tree	55,54	64,67	9,13
**: p<0,05; **: p<0,01*			

As seen in Table 2, there is only one significant mean difference. The mean score, expressed as a percentage, of the subjects in group 1 and thus with the raw US set, score a mean of 13,01% points significantly higher than the subjects in group 2. In other words, there is a significant decrease in the mean of the score of 20,11% in group 2, compared to group 1. Part 2 and the exercise on the RT show no significant difference in means. The expectation that the QUS-compliant US set would improve the overall scores of the exercises that are executed by the subjects is not confirmed. On the contrary, subjects from group 1 score higher on the exercises in part 1. Although, the means of the scores from the exercises for part 2 and the RT are higher in group 2, they are not significant. A plausible explanation for the mean difference in the exercises of part 1 is rather hard to find while similar, but improved, US sets are used in group 2. It might be possible that the effect of the QUS framework, that changed some of the US, and the selection of a few different US influenced the ability of the novice modelers to identify the concepts in part 1. To test whether the mean differences are significant, an independent t-test was conducted for part 1 and 2 (Appendix N). The means for the modeling exercises are tested according to the Kruskall-Wallis test, because in group 1 the variable is not normally distributed (Appendix L).

Besides the overall score, the scores of the separate exercises in part 1 and 2 have also been analyzed. Table 3 shows the differences of the separate exercises between group 1 and group 2 and indicates whether they are significant. Again, percentages are used to ensure consistent comparisons. The mean differences are tested by a Kruskall-Wallis test (Appendix M & L). The mean difference of the scores in identifying Themes and Epics in part 1 between both groups is significant. There is a decrease of 32,14% in the mean score from group 1 to group 2 in identifying Themes. The mean score of the identification of Epics in group 1, is significantly higher than in group 2. These differences explain the mean difference of the overall score in part 1. Another explanation could be that the QUS-compliant set had a negative impact on the subjects' abilities to identify Epics and Themes from a short set of US. Although the mean scores' differences are not significant, Table 3 shows that the mean scores of identifying Themes and

Epics are higher in group 2 from a between-group point of view, but especially from a within-group point of view. From these results, a new hypothesis can be raised: *the subjects' ability to identify Themes and Epics within a high-quality set of US improves while using a RT to identify them*. The hypothesis that a QUS-compliant set will improve the identification of Epics, Themes, NFRs and missing requirements is rejected in both the cases before and after the use of the RT.

Table 3. Separate scores of the exercises in part1 and part2.

Variables	Mean group 1	Mean group 2	Mean difference (% points)
Percentage of score NFR part 1	74,26	70,00	4,26
Percentage of score Themes part 1	82,35	55,88	26,47*
Percentage of score Epic part 1	73,53	20,59	52,94**
Percentage of score MR part 1	23,53	26,47	2,94
Percentage of score NFR part 2	73,53	73,53	0,00
Percentage of score Themes part 2	47,84	62,75	14,91
Percentage of score Epic part 2	41,18	64,71	23,53
Percentage of score MR part 2	35,29	29,41	5,88
**: p<0,05; **: p<0,01*			

Table 4 shows the means of the scores (in points, not as percentages) for the separate modeled elements of the RT. To clarify the figures, the maximum amount of points that could be given to a subject for each variable is indicated. According to the table, the subjects could best identify the coarse-grained functionalities in both group 1 and 2. The average score of the subjects was also high for modeling a consistent RT. That finding can be linked to the research of Wautelet et al. [11] which concluded that most of the subjects could create an acceptable graphical US model. The subjects scored the least points in identifying the missing links, an error that also frequently occurred in the mentioned study. When looking at the mean differences, there are three values that show a significant difference. The mean score for modeling the tasks, the capabilities and the links is significantly higher in group 2. This implies that some of the expectations are partially confirmed. In both exercises the same US set was used, the only difference was the interference of the QUS framework to improve the quality of the US set. A plausible explanation for the significant difference might be that a US set of better quality (i.e., improved by the QUS framework) helps the modeler to identify some elements of the RT better, specifically tasks, capabilities and links. This could be an interesting finding, while Wautelet et al. [11] mentioned a lot of modeling errors concerning the capability

element. The interference of the QUS framework could be a possible solution to easily identify atomicity in functional elements.

A non-parametric Kruskall-Wallis test (Appendix M) was used to compare the means of all the variables in Table 4, except for the score attributed to identifying the links. For the latter, an independent t-test (Appendix N) was executed because the normality condition was met (Appendix L).

Table 4. Comparing scores on the elements of the Rationale Tree.

Variables	Mean group 1	Mean group 2	Mean difference
Score modelled 3 coarse-grained functionalities in Tree (3p)	2,4118	2,5294	0,1176
Score modelled 3 hard-goals in Tree (3p)	1,2353	1,5882	0,3529
Score modelled 2 soft-goals in Tree (2p)	1,4706	1,4706	0,00
Score modelled 4 tasks in Tree (2p)	0,8824	1,2353	0,3529*
Score modelled 2 capabilities in Tree (1p)	0,6176	0,8824	0,2648*
Score modelled 8 links in Tree (4p)	1,6471	2,3765	0,7294*
Score modelled a consistent Tree (1p)	0,8824	0,7941	0,0883
Score identifying missing links (1p)	0,2941	0,1176	0,1765
**: p<0,05; **: p<0,01*			

The **perceptions** of respondents have also been analyzed. Due to a lack of space and because it is not fundamental for the overall understanding of the paper, it has been placed in Appendix M.

4.3 A within-group comparison: analyzing the impact of the Rationale Tree

In this section, a within-group analysis is made. Like in the previous section, the different scores will be compared by testing whether there is a significant difference, but the means of the exercises from the different parts are here compared in both groups separately. The main goal is to evaluate whether the use of the RT improves the ability of the subject to identify different concepts and to test whether the impact of the RT improves while using a US set of higher quality. Within this section, the new conducted hypothesis from Section 4.2 will be tested.

Analyzing the scores First, the overall scores of the exercises in both parts are compared. Figure 1 depicts the overall mean scores, as percentages, of the exercises from part 1 and 2 for both groups. The figure depicts the previously identified significant difference in the exercises of part 1 between both groups. Within group 1 (0 in the chart) and group 2 (1 in the chart), the paired t-test is used to test whether there was a significant difference between both parts. The t-test shows there is a significant ($p<0,01$)

difference in group 1 between the exercises of part 1 and 2. The tests show there is no significant difference in group 2. With respect to the previous tests, analyzed in Section 4.2; it is clear that differences exist in the overall score of the exercises in part 1, both within and between the groups. An explanation for the within-group difference might be that the RT does not help the test subjects to identify the concepts when using a US set of lower quality. Besides that, part 2 introduces something totally new to the test subjects, the RT, that could also have an influence on the ability of the modelers to make the exercises. Another possible explanation of the difference could be the usage of different US sets in both parts. Additionally, the US set in part 1 of group 1 was slightly different from the US set in part 1 of group 2.

A second within-group comparison is done by analyzing the mean differences in the scores of the separate exercises. Table 5 explains the significant difference between the overall scores of the exercises in part 1 and 2. According to a non-parametric Wilcoxon Signed Ranks test, the mean differences of the scores on the exercises regarding Themes and Epics are significantly different. The data shows that group 1 better identified Themes and Epics in part 1. That finding also aligns with the significant difference in the means of the scores on identifying Themes and Epics between both groups. With respect to the possible other explanations for the significant difference, the explanation in the previous paragraph could be refined into the following: *the RT does not help the test subjects to identify Themes and Epics when using a US set of lower quality*.

Table 5. Comparing separate exercises group 1.

Variable	Mean part 1	Mean part 2	Mean difference
Percentage of score on exercise Theme	82,35	47,84	34,51**
Percentage of score on exercise Epic	73,53	41,18	32,35*
**: p<0,05; **: p<0,01*			

In the Table 6, the same comparison is made but now from the point of view of group 2. As in Table 5, only the relevant variables are depicted. The data shows that test subjects can better identify Epics after using the RT. The difference is significant. A plausible explanation might be that the RT helps identifying Epics when using a high-quality set of US. The new hypothesis can thus be partially accepted (only concerning Epics).

Table 6. Comparing separate exercises group 2.

Variable	Mean part 1	Mean part 2	Mean difference
Percentage of score on exercise Epics	20,59	64,71	44,12**
**: p<0,05; **: p<0,01*			

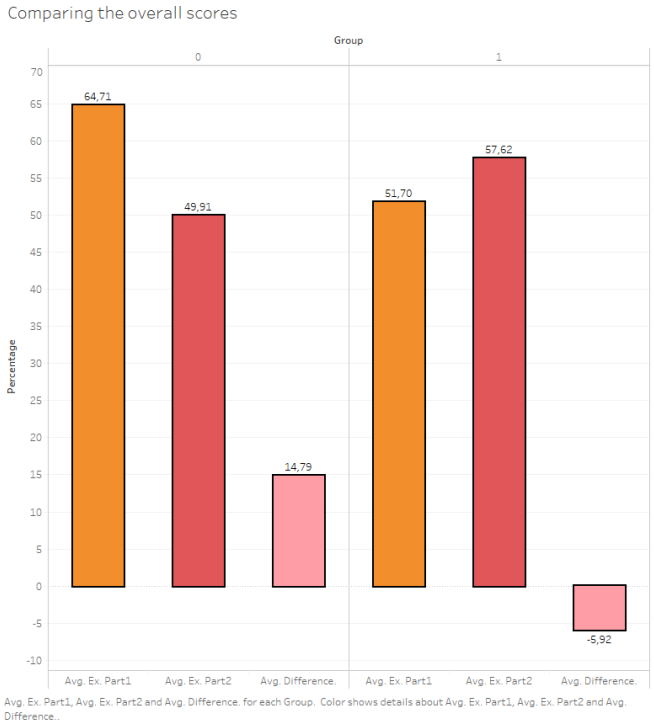

Fig. 1. Comparing the overall scores of the two parts.

5 Threats to Validity, Future Work and Limitations

The first and main threat to validity comes from the "distance" between the raw and QUS-compliant US sets. We have selected two sets of US that have been improved using the QUS framework without having a quantitative evaluation of the distance between the two sets (it is up to the reader to evaluate this distance by tracing the revision procedure and/or reading the initial and QUS-compliant sets). It could be that (raw) US sets of various initial qualities do exist within real-life US sets and that the QUS application will bring more value to initial US sets with lower quality. This would have a direct impact on the ability of the modeler to understand the software problem, to identify functions, their abstraction and complementarity as well as elements like NFRs, Epics, Themes and missing requirements. We need to establish a way to measure/quantify the distance between the raw and QUS-compliant US sets and reproduce the experience with sets having different distances to better understand this. Another threat comes from the quoting system itself. The latter has been built through an analysis of default solutions and a moving golden standard with the aim to define the criteria making the representations relevant and of high quality. While we have included all of the possibilities we found and justified the importance of the criteria we used, other solutions could perhaps have been included.

We also point out two limitations. First, the experiment was only executed by students that, despite some different educational backgrounds, all studied Business Administration. In future research, it would be interesting to compare the ability of different sample groups like agile/requirements specialists, business analysts or other students with a different background (e.g., computer science students). A second limitation concerns the limited amount of information that was given to the subjects. Despite the previously mentioned introduction about US and the information given about the different concepts related to US sets, the amount of information was still limited for students without any previous knowledge about the concepts. Also, the presentation of the RT and its concepts was kept as limited as possible so subjects could execute the experiment within the time frame (approximately 2 hours). An introduction and explanation about the unified model for US modeling, for example, was not given to the subjects, although knowledge about that would have been useful.

6 Conclusion and Future Work

After describing the data and creating factors, two types of comparisons were made. A between-group comparison and a within-group comparison were indeed conducted to measure the impact of both the QUS framework and the RT. Some significant differences were found from which the following main conclusions could be drawn. Applying a high-quality US set compared to a US set of lower quality did not improve the test subjects' ability to identify the US related concepts (themes, epics, NFRs or even missing requirements) that were tested in the exercises, both with and without the use of the RT. A possible explanation for the rejection of the hypothesis was that group 2 received a slightly different US set than group 1 in the first part. The improved US set could have been experienced as more difficult for the novice modelers in group 2. The non-significant differences in part 2 between both groups might be explained by the interference of a new framework that the novice modelers did not know. Neither did the interference of the QUS framework improve the overall scores of the exercises compared to the US set of lower quality. Overall, we thus cannot conclude that the effect of the QUS framework, compared to a US set of lower quality, had any benefits to understand the problem/solution domain of the real-life case. A finding that did confirm an expectation was that the QUS-compliant US set improved the ability of the test subjects to identify and model some parts of the RT better, specifically Tasks, Capabilities, and links. This could be due to the fact that the QUS-compliant US set is more consistent and less overlapping than the raw one so helping the modeler to better separate and structure the elements present in US. While analyzing the data, a new hypothesis could be developed. According to some clear differences in means, there was expected that a QUS-compliant US set could improve the test subjects' ability to identify Themes and Epics with the use of the RT compared to identifying the same concepts without using the RT. That expectation was only partially confirmed because there was only a significant difference regarding the identifications of Epics. Even if building a RT out of a US set of a higher quality level does not impact the ability of test subjects to identify Themes, Epics or missing requirements, we can conclude that building the RT from a QUS-compliant US set improves the ability of the novice modeler to identify Epics. By

helping in this identification, a RT built out of a QUS-compliant US set improves the ability to understand the problem/solution domain in a real-life case.

7 Acknowledgement

The authors would like to thank Fabiano Dalpiaz for having evaluated the raw US set with the AQUSA tool and Duje Delic for his involvement in the experiment.

References

1. 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, 2013. pp. 115–124. IEEE Computer Society (2013)
2. Cohn, M.: Succeeding with Agile: Software Development Using Scrum. Addison-Wesley Professional, 1st edn. (2009)
3. Liskin, O., Pham, R., Kiesling, S., Schneider, K.: Why we need a granularity concept for user stories. In: Proceedings of XP'14, Rome. LNBIP, vol. 179, pp. 110–125. Springer (2014)
4. Lucassen, G., Dalpiaz, F., van der Werf, J.M.E., Brinkkemper, S.: Improving agile requirements: the quality user story framework and tool. Req. Eng. **21**(3), 383–403 (2016)
5. Lucassen, G., Dalpiaz, F., van der Werf, J.M.E., Brinkkemper, S.: Improving user story practice with the grimm method: A multiple case study in the software industry. In: Int. Working Conference on Req. Eng.: Foundation for Software Quality. pp. 235–252. Springer (2017)
6. Taibi, D., Lenarduzziand, V., Janes, A., Liukkunen, K., Ahmad, M.O.: Comparing requirements decomposition within the scrum, scrum with kanban, xp, and banana development processes. In: Springer (ed.) Proceedings of XP 2017, LNBIP 283 (2017)
7. Trkman, M., Mendling, J., Krisper, M.: Using business process models to better understand the dependencies among user stories. Information & Software Technology **71**, 58–76 (2016)
8. Wautelet, Y., Heng, S., Kiv, S., Kolp, M.: User-story driven development of multi-agent systems: A process fragment for agile methods. Computer Languages, Systems & Structures **50**, 159–176 (2017)
9. Wautelet, Y., Heng, S., Kolp, M., Mirbel, I.: Unifying and Extending User Story Models. In: CAiSE 2014, Thessaloniki, Greece. Proc. LNCS, vol. 8484, pp. 211–225. Springer (2014)
10. 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, June 1-3, 2016. pp. 477–488 (2016)
11. Wautelet, Y., Velghe, M., Heng, S., Poelmans, S., Kolp, M.: On modelers ability to build a visual diagram from a user story set: a goal-oriented approach. In: Int. Working Conference on Requirements Engineering: Foundation for Software Quality. pp. 209–226. Springer (2018)
12. Yu, E., Giorgini, P., Maiden, N., Mylopoulos, J.: Social Modeling for Requirements Engineering. MIT Press (2011)