International Scholarly Research Network
ISRN Software Engineering
Volume 2012, Article ID 795282, 10 pages
doi:10.5402/2012/795282

Research Article

Evaluating a Taxonomy for Mobility Requirements by a Controlled Experiment

Sundar Gopalakrishnan, Peter Karpati, and Guttorm Sindre

Department of Computer and Information Science, Norwegian University of Science and Technology, IDI/NTNU, NO-7491 Trondheim, Norway

Correspondence should be addressed to Guttorm Sindre, guttors@idi.ntnu.no

Received 22 August 2011; Accepted 4 October 2011

Academic Editor: U. K. Wiil

Requirements taxonomies have been found useful in software requirements elicitation and specification, both for educational purposes and in practical usage, for instance, as checklists to ensure that important categories of requirements are not forgotten, and for guidance on how to write various types of requirements. While mobile information systems are becoming increasingly important, traditional requirements taxonomies do not have any category for mobility requirements. This paper reports on a controlled experiment where two groups of students both got the same excerpts of the well-known Volere requirements taxonomy, but for one treatment group the tutorial material was also extended with additional material on mobility requirements as a requirements category in its own right. Using the provided taxonomy material for guidance, the students were asked to write requirements for a system presented in a natural language case description; afterwards their output was analyzed to score the number and quality of requirements found by each student. The main finding was that the students using the extended taxonomy also found more requirements, but there was no significant difference in the quality of requirements between the two groups.

1. Introduction

Requirements engineering is a crucial activity in the development of large software systems, bridging between the customer's needs and what can realistically be implemented in software. Requirements taxonomies have been found useful for this activity, both for educational purposes and in practical usage, for instance, as checklists to ensure that important categories of requirements are not forgotten, and for guidance on how to write various types of requirements. One well-known generic taxonomy for software requirements is the Volere taxonomy [1], which discusses functional and data requirements in one section, and then a number of different categories of quality requirements in the next section. On the top level, these include look and feel requirements, usability, performance, operational requirements, maintainability and support requirements, security, cultural and political requirements, and legal requirements. Most of these are further broken down into various subcategories, for instance security into access, integrity, privacy, audit, and immunity requirements. Another and more detailed taxonomy with a larger hierarchy of categories and subcategories of quality requirements has been proposed by Firesmith [2]. In addition, there are requirements taxonomies limiting themselves to exploring some quality factors in more detail, for instance for safety [3], security [4], privacy [5], sustainability [6], or trust-related requirements [7], or for specific technologies, such as wireless sensor network applications [8].

Recently, mobile devices have seen a lot of improvements in computing power, user interfaces (e.g., smart phone touch screens), and bandwidth. This has made it possible to perform a wide range of information processing tasks on the go that previously had to be performed in the office. This has increased the importance of mobile information systems. Many mainstream requirements techniques were however developed before mobile information systems were in widespread use, and this also applies to the general requirements taxonomies [1, 2] mentioned above. To better

understand the area of mobility-related requirements, we proposed a taxonomy for this in some previous publications [9, 10], and then in [11] we elaborated parts of this taxonomy further into requirements boilerplates (templates) in the style of [12]. However, none of these previous publications have evaluated empirically whether a taxonomy of mobility-related requirements would really help finding more requirements or specify them with better quality. This paper is the first step in such an evaluation, performing an experiment whether people will find more or better mobility requirements if provided with a taxonomy for that category of requirements than they would without it.

The research questions for the experiment are as follows:

RQ1. Does the taxonomy of mobility-related requirements help people finding more requirements?

RQ2. Does the taxonomy of mobility-related requirements help people write requirements of better quality? In the context of this experiment, quality was defined in terms of relevance, testability, and clarity of the requirements.

The rest of the paper is structured as follows: Section 2 deals with background and related work in requirements taxonomies. Section 3 discusses research methodology. Section 4 presents the results, and Section 5 discusses threats to validity. Section 6 concludes the paper.

2. Background and Related Work

Before getting into the experiment part, a brief introduction of the Volere taxonomy and the proposed Mobility extensions will be useful. Volere [1] is well-known in the requirements field. In addition to the requirements taxonomy, which categorizes requirements as functional, data, or quality requirements (Figure 1) and breaks quality requirements into further subcategories, it also provides a requirements specification template that may be used with Requisite, DOORS, Caliber RM, IRQA, and other popular tools.

Volere does not have any particular category for mobility requirements, but it does include some other categories which are clearly related to this. In particular, the category "physical environment" may cover requirements that the system shall be used in, for example, wet, noisy, or warm settings. Also, categories like reliability, availability, and robustness—although generally relevant for all kinds of systems—will have special challenges for mobile systems, for example, because of poor or lost network coverage and various physical impacts on the equipment. Some of the requirements that might be captured as mobility requirements according to our taxonomy in [9–11] might therefore be represented in these categories in the Volere taxonomy.

Our proposed taxonomy for mobility-related requirements [10] is much inspired by a similar taxonomy by Firesmith [4] for security-related requirements. Both Firesmith's taxonomy and our mobility taxonomy combine requirements from a challenge part and an achievement part, that is, given some challenge (X) the system should achieve (Y). For security, the challenge might be a certain type of attack (e.g., an intrusion attempt to the system) and the achievement level something the system should be able to do in this case (e.g., detecting or preventing the attack). For mobility, the challenge might instead be that the user needs to travel, for example, at a certain speed, and still the system needs to sustain a network connection, give good navigation advice, and so forth. In more detail, the proposed taxonomy includes four categories of requirements:

(i) (pure) mobility requirements: specifying some level of mobility, without indicating design solutions, that is, specifying mobility challenging factors and mobility achievement levels. Examples for mobility challenge factors may be

  (a) the speed of movement needed (the larger the speed, the more difficult it might be to support the movement or provide nondegraded service while moving),

  (b) the area/range of movement (the larger the area, the more difficult).

  Examples for mobility achievement levels may be

  (a) ability to (actively) move. This could be particularly relevant for embedded systems, for example, where a software application is running an engine and steering system, but less relevant for enterprise information systems, which is our main concern,

  (b) ability to facilitate movement, for example, real-time positioning, mapping, and navigational services,

  (c) ability to provide a certain level of service (e.g., no degradation from the normal situation) in spite of the challenging usage context,

(ii) mobility-system requirements: requirements associated with subsystems whose purpose is to support mobility. It obviously depends on whether the system needs to be in a particular network or not. Examples may be positioning system, network scanning and acquisition system, service scanning and acquisition system;

(iii) mobility data requirements: data requirements werer associated with subsystems which support data and mobility. Examples may be data the system needs to store mobile devices in use, their contact details, storage capacity, position data, tracking data, and so forth;

(iv) mobility constraints: design decisions related to mobility that have been lifted to the requirements level. Examples may be decision to use one specific standard for mobile communication, decision to use one specific type of mobile equipment, or equipment compatible to that—given that the customer really *requires* this particular decision.

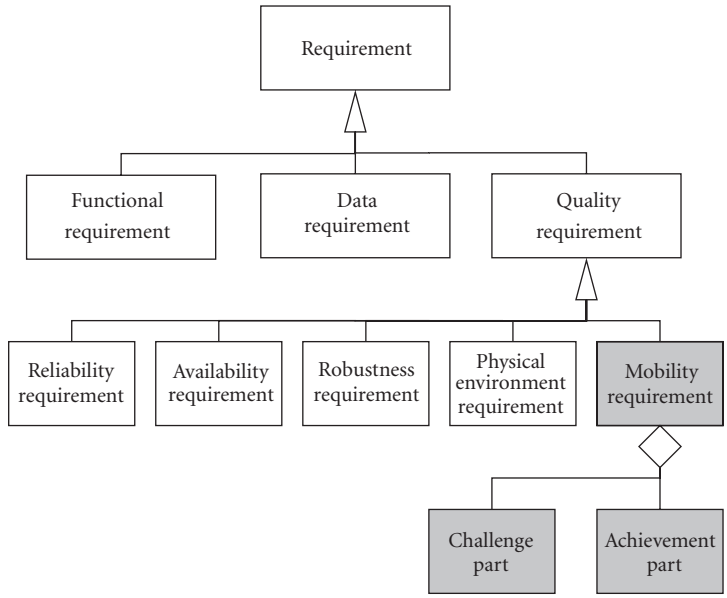

Figure 1: Volere taxonomy (white), including only level 3 nodes relevant for the experiment, and extensions for mobility (grey).

Both the Volere taxonomy and our mobility taxonomy are too big to be explored in their entirety in just one single experiment. The most interesting part of our taxonomy if considered to have mobility requirements as a category in their own right in a large, more general taxonomy, is the pure mobility requirements (first black bullet item in the list above, with white subbullets). So, as a starting point we took just a simplified version of our taxonomy, focusing on this part, and likewise just an excerpt of the Volere taxonomy that would be most closely related to this, namely, the subcategories of reliability, availability, robustness, and physical environment. The requirements taxonomies that were used in the experiment are thus indicated in the UML class diagram in Figure 1. The white parts are the excerpts of the Volere taxonomy, for which there was a tutorial about level 3 subcategories given to the experiment participants. The grey parts is the extra category of mobility requirements, for which a similar tutorial was given. Our treatment group thus got the entire taxonomy (tutorial for white + grey parts) while the control group got only the white parts. Both groups were recommended to make requirements as precise and quantify them where possible, and effort was made to provide a similar amount and style of examples and explanation for the mobility category as for the other ones.

For the mobility category, the tutorial explained that such requirements might typically contain a challenge part and an achievement part, as indicated by the aggregation relationship at the bottom of Figure 1. So, the tutorial recommended such requirements to be written "The application shall ⟨*achievement level X*⟩ when ⟨*challenge Y*⟩". Examples of such requirements were provided in the tutorial: (1) "The application shall provide non-degraded service when users are moving at up to 100 mph along major highways." and (2) "The application shall be able to resume operation with the session status intact when network connection has been temporarily lost for up to 10 minutes." Of course, one does not need to have mobility as a category in its own right in a taxonomy to be able to come up with requirements like these, for participants using only the white part of the taxonomy, they could be thought of as, for example, availability or reliability requirements. The assumption for the experiment, thus, is not that a category for mobility requirements will enable people to write requirements that they could not possibly have written before, rather it is that they might be helped in eliciting and specifying such requirements more effectively.

To our knowledge, there is no other paper reporting an experiment comparing two software requirements taxonomies for the purpose of elicitation and specification of requirements. Reference [13] reports on an experiment assessing the usefulness of a taxonomy for design requirements, but this was in the area of concurrent engineering. Similarly [14] discusses several approaches for preparing requirements taxonomies in the area of engineering product design. Within the software field, Volere was tried out for specifying a mobile learning support system in [15], one of the experiences being that more flexibility might be useful in the requirements categorization than what was offered by the fixed Volere categories.

Reference [16] discusses on a more general level the challenge of categorizing nonfunctional requirements. References [5, 6] explore another new categories of requirements, namely, privacy and sustainability requirements, but validated through case studies rather than experiments.

3. Research Method

Our research goal in this paper is to find out whether a taxonomy for mobility-related requirements will help people find more and better such requirements given a requirements

specification task. Typical ways to perform such evaluations are either case studies or controlled experiments, between which there are some obvious trade-offs, especially concerning realism (in favor of case studies) versus control (in favor of experiments) [17]. A case study can feature large tasks, possibly entire industrial projects applying the techniques we want to evaluate, but the lack of control means that it is difficult to make comparisons. For example, even if our project X which applies our proposed requirements taxonomy turns out successful, it will be hard to know whether the success was really caused by the taxonomy or by something completely different (e.g., clever analysts, customer who was clear about needs, easy project task, and other RE techniques or tools that were also used in the project, etc.). Also it would be hard to know whether the project would have been any less successful if applying another requirements taxonomy. Such comparisons will be much easier with a controlled experiment, where one group of participants can use the taxonomy for a certain RE task while another group uses something else. On the other hand, controlled experiments must necessarily be of limited duration, which means that the RE task to be solved must be fairly small and simple, thus not fully representative of the complexity facing developers in real industrial projects. Since our research question for this work was clearly of a comparative nature, we decided to go for an experimental research approach, leaving industrial case studies for future work. This also makes sense from the perspective of time-consumption. Although controlled experiments demand quite a lot of work, case studies usually demand even more. If experiments were to indicate that our proposed technique does not have any advantages, it might therefore not be worth spending time on case studies at all. And on the other hand, positive results from experiments could more easily motivate for case studies and potentially make industry more interested in participating in such a study.

3.1. Experiment Design. For practical reasons, it was decided to run the experiment with student participants. While industry participants might have been better in terms of being more representative of the state-of-practice, we lacked the funds to hire practitioners by the hour to participate in such an experiment. Moreover, since the experiment was about a new requirements taxonomy, which the participants had to learn and then use in the time frame of the experiment, the difference between students and practitioners might be smaller than it would have been if, for example, comparing two RE techniques well known in industry already. That is, in our experiments, practitioners would, in the same manner as students, have had to start by reading about the proposed taxonomy and then try to apply it.

A vital challenge for the experiment design would be exactly what to compare. Our main alternatives would be

(1) the treatment group get our mobility requirements taxonomy, the control group get nothing, thus using an ad hoc technique,

(2) the treatment group get an existing requirements taxonomy extended with our mobility requirements taxonomy, the control group get only the existing requirements taxonomy without any extension.

The perceived problem here was that alternative 1 would be too biased in favor of the new taxonomy, since—through the tutorial about this taxonomy—the students would get a lot of hints about possible types of requirements and how they could be written, while the control group would get nothing. This could have worked if the participants were experienced RE practitioners who could then use instead their state-of-practice technique, but with students lacking such a technique, the result for the control group might easily be very limited or poor output. We therefore chose alternative 2, both groups getting the same excerpt of the Volere taxonomy, and the treatment group additionally getting one extra category of requirements in their experiment tutorial on mobility requirements. Thus, both groups would get some hints on types and styles of requirements, only the treatment group would get a little more. This experiment could more easily go both ways, the treatment group could gain from the extra material if the mobility-related taxonomy turned out useful for identifying some requirements, or alternatively—if this taxonomy turned out confusing or not fitting the other material, or the treatment group were thus overloaded with information compared with the control group—it could also be detrimental to their performance.

Another challenge was that we could not include the entire Volere taxonomy and entire taxonomy of mobility-related requirements in the experiment, as this would have caused the tutorial material to be way too long to fit into the time frame of a controlled experiment. Hence, as mentioned before, from the taxonomy of mobility-related requirements, we only included the so-called pure mobility requirements. Similarly, from Volere we only included the categories most closely related to mobility, namely 12d reliability and availability, 12e robustness and fault tolerance, and 13a Expected Physical Environment.

54 students were recruited from a second year computer science class to take part in the experiment. The participants were randomly divided into 2 groups, the treatment group using excerpts from the Volere taxonomy extended with a part about mobility requirements, and the control group getting only the excerpts from Volere.

The participants performed the following tasks during the experiment:

(1) Answering a preexperiment questionnaire. The purpose of the preexperiment questionnaire was to investigate the participants' prior knowledge of related topics like mobility, requirements specification, and so forth, which can be used to control for any accidental group selection bias in spite of random selection (e.g., one group accidentally containing people with much more relevant experience). Questions investigated previous knowledge on eliciting requirements and specification, IT work experience, and familiarity with portable mobile devices, in

total 8 questions that were to be answered within 5 minutes.

(2) Reading a tutorial about the requirements taxonomy to be used during the experiment, either a pure Volere excerpt for the control group or the same Volere excerpt extended with a part about mobility requirements for the treatment group.

(3) Specifying requirements for a mobile application for airline check-in, for which a case description was provided in natural language. From the case description, which was simple running prose, the students were asked to write suggestions for natural language "The system shall…" requirements, as many relevant requirements as they could come up with during the allotted time. The precise task stated that a web-based system usable on stationary office PCs already existed for the system in question, so that their task was not to propose requirements for the system in general, only to propose additional requirements that would come up when the system owner now wanted to make the system usable also on mobile devices.

The motivation for focusing exclusively on mobile requirements in item 3 is of course that it is such requirements the mobility taxonomy purports to support. We have no assumption that the mobility taxonomy would be of any help if specifying requirements for a traditional stationary information system, so experimenting with the taxonomy in such a context would not be very useful.

3.2. Variables and Hypotheses. From investigating the subjects' task performance, we are interested in finding how many requirements each would identify and the quality of these requirements. It will also be interesting to look not only at the total number of requirements identified, but also at different types of requirements. Since the only difference between the two treatments is that one tutorial contained some extra material about mobility requirements that the other one did not contain, the main effect—if any—would be assumed to materialize for this particular type of requirement. Moreover, since mobility requirement is considered a subtype of quality requirement, there could again be some effect on these. Hence we divide requirements in 3 disjoint categories: mobility requirements, other quality requirements (i.e., which are not mobility requirements), and other requirements (i.e., which are not quality requirements). In some cases, it might also be interesting to look at unions of these, for example, all requirements taken together, or all quality requirements taken together (i.e., including mobility requirements).

As for quality of requirements, there are many possible measures for this. For instance, Knauss and El Boustani [18] identify a long list of possible quality measures for requirements and requirements specifications. However, not all these are equally relevant for our purposes. First of all, some are more related to the quality of the requirements specification as a whole or the relationship between requirements—or between requirements and other documents, for example, completeness, traceability, and consistency. Since one would not expect to achieve completeness in a brief experiment where students are given limited time to find requirements, such properties are not so relevant here. Moreover, some properties are about typos and grammar, proper numbering, and similar syntactic matters; again, this would be highly relevant in an industrial setting but hard to enforce in a brief experiment with students. Again, other attributes suggested in [18] are about user interface descriptions, metrics, and other issues that went way beyond the scope of our experiment. Also, our goal is not to evaluate every possible quality attribute that we might possibly think of, which would be prohibitively expensive in terms of time spent on data analysis, but to look at some generally important and representative quality attributes that can be considered for each single requirement—just to check that, for example, a treatment that shows an advantage in finding more requirements does not lose its advantage, by at the same time, delivering poorer quality of requirements. Hence, the following three quality attributes for requirements were chosen in this experiment:

(i) relevance: the degree to which the proposed requirement is relevant to the presented case or not, ranging from 3 (obviously relevant) to 0 (not at all relevant). Relevance is pointed out as an important property of requirements in [19]. A related term, used in [18] and many other sources, would be "correctness". However, this assumes a real problem with a correct solution or a stakeholder whose requirement is to be correctly represented. In our experiment, there was no customer present, only a fairly vague case description upon which the students were asked to suggest requirements. Hence, correctness cannot be stringently evaluated and relevance therefore seems a better choice of criterion;

(ii) clarity: the degree to which the requirement's intent is clearly understandable to the reader. This is also pointed out as an essential property of requirements in a survey reported in [20]. Another related term is "lack of ambiguity", for instance, in [19];

(iii) testability: the degree to which it would be possible to write a test which clearly demonstrates whether the requirement has been satisfied or not. For quality requirements this will typically mean some kind of quantification. Again, this is pointed out as essential in [20], and in several other sources, for example, as "verifiability" in [19].

All in all, this gives us a large number of variables to be measured. To save space later in the paper, each variable is named with three letters. The first letter indicates the measure (number, relevance, clarity, etc.), the second the category of requirements that the measure applies to (mobility requirements, other quality requirements, etc.), and the third letter indicates the group of experimental subjects that the measure applies to (treatment group or

control group). Examples of variables are the following:

(i) NMT, NMC, the number (N) of mobility requirements, (M) found by the treatment group (T) or control group (C), respectively;

(ii) NXT, NXC, the number (N) of quality requirements except mobility, (X) found by the treatment group (T) or control group (C), respectively;

(iii) NQT, NQC, the number (N) of quality requirements (Q = M + X) found by the treatment group (T) or control group (C), respectively;

(iv) NOT, NOC, the number (N) of other requirements (O, that is, not quality requirements) found by the treatment group (T) or control group (C), respectively;

(v) NAT, NAC, the number (N) of all requirements (A = Q + O) found by the treatment group (T) or control group (C), respectively;

(vi) RMC, RMT, the relevance (R) of mobility requirements, (M) found by the treatment group (T) or control group (C), respectively;

(vii) CMT, CMC, the clarity (C) of mobility requirements, (M) found by the treatment group (T) or control group (C), respectively;

(viii) TMT, TMC, the testability (T) of mobility requirements, (M) found by the treatment group (T) or control group (C), respectively.

(ix) QMT, QMV, the quality (Q) of mobility requirements, (M) found by the treatment group (T) or control group (C), respectively. The quality is calculated as the average of the three quality criteria used, namely relevance, clarity, and testability.

Similar variables would then result for the quality of other quality requirements (e.g., RXT,…QXC), quality requirements, other requirements, and all requirements—for the sake of brevity these are not listed in detail since it is assumed to be fairly obvious what all these variables would be.

For the posttask questionnaire, we have some questions related to perceived ease of use (PEOU), some to perceived usefulness (PU), and some to intention to use (ITU). Hence it makes sense to measure 3 variables for each treatment group:

(i) PEOU_T, PEOU_C, the response to the PEOU questions by the treatment group (T) or control group (C), respectively;

(ii) PU_T, PU_C, the response to the PU questions by the treatment group (T) or control group (C), respectively;

(iii) ITU_T, ITU_C, the response to the ITU questions by the treatment group (T) or control group (C), respectively;

(iv) P_T, P_C, the overall preference for the taxonomy (i.e., average of PEOU, PU, and ITU) for the treatment (T) and control (C) groups, respectively.

For each of these variable pairs, there will then be a null hypothesis stating that there is no difference between the two variables, and an alternative hypothesis stating that there is a difference, for example,

(i) $H_{0,NM}$: there is no difference in the number of mobility requirements found by the two groups (i.e., NMT = NMC);

(ii) $H_{1,NM}$: there is a difference in the number of mobility requirements found by the two groups (i.e., NMT $\neq$ NMC);

(iii) $H_{0,TO}$: there is no difference in the testability of other (non-quality) requirements found by the two groups (i.e., TOT = TOC);

(iv) $H_{1,TO}$: there is a difference in the testability of other (non-quality) requirements found by the two groups (i.e., TOT $\neq$ TOC);

(v) $H_{0,PEOU}$: there is no difference in the perceived ease of use of taxonomies reported by the two groups (i.e., PEOU_T = PEOU_C);

(vi) $H_{1,PEOU}$: there is a difference in the perceived ease of use of taxonomies reported by the two groups (i.e., PEOU_T $\neq$ PEOU_C).

These are just 3 out of a total of 24 pairs of hypotheses, but again we leave out the rest for space reasons, as it would be fairly monotonous to list them all. It can be noted that all the alternative hypotheses are formulated without indicating any expected direction for the difference. Of course, one might think that, for example, the number or quality of mobility requirements should increase for the treatment group, which gets the mobility taxonomy. However, since the taxonomy has not been tried in practice before, it could also happen that the extra taxonomy was only confusing and actually led to fewer requirements or poorer quality. It therefore seemed most appropriate to use hypotheses without any particular assumptions about the outcome.

4. Experiment Results and Analysis

4.1. Analysis of the Posttask Questionnaire. On average, the participants' response to the questionnaire was slightly more positive for the treatment group than for the control group, the total average response for all questions being 3.5 versus 3.3, respectively. Since this was scored on a 5-point Likert scale where 3.0 would be the average, it seems that the respondents have been slightly positive but not really enthusiastic about any of the taxonomies. Table 1 shows the averages for questions related to Perceived Ease of Use, Perceived Usefulness, and Intention to Use, respectively. As the table indicates, there were no significant differences. Even looking at single questions, the question with the biggest response difference (Q5, mean response being 3.8 versus 3.3) did not show significant results ($P = 0.09$ in a simple t-test). Hence the null hypothesis could not be rejected in this case, so the natural conclusion is that there was no difference in participants' opinion about the two taxonomies.

Table 1: Results for the posttask questionnaire.

Variable	Treatment group		Control group			
	Mean	St.dev.	Mean	St.dev.	Diff	Sign.?
PEOU	3.74	0.62	3.69	0.69	0.05	No
PU	3.64	0.71	3.35	0.83	0.29	No
ITU	3.14	0.72	2.85	0.93	0.29	No

4.2. Analysis of Task Performance. For the task performance, we were investigating whether one group would find more requirements than the other group, and whether there would be any difference in the quality of the requirements. Three scores were given for the quality of each requirement, namely its relevance, clarity, and testability. Moreover, requirements were classified either as "mobility requirement", "other quality requirement" (except mobility), and "other requirement" (except quality).

Table 2 shows the results concerning the number of requirements, using the three-letter naming scheme for variables as explained before, except the third letter (T or C) has been replaced by a question mark since each row shows data both for the treatment group and control group. As indicated, the treatment group found notably more mobility requirements than the control group, the difference being 1.09, giving an effect size of 0.68, which is a moderate effect [21]. They also found slightly more other quality requirements (i.e., quality requirements that were not classified as mobility requirements) but on the other hand they found slightly fewer other requirements (i.e., requirements that were not classified as quality requirements, for instance functional or data requirements); however, these effects were only small. Because of the notable advantage for mobility requirements, the number of quality requirements (including mobility, that is, NQ) as well as the total number of requirements (including quality and mobility, that is, NA) was also bigger for the treatment group, with effect sizes of 0.55 and 0.40, respectively, which could be classified as small to moderate effects.

As the data sets passed an Anderson-Darling test for normality, a simple t-test was used to compare the two groups. Only two of the differences were significant, namely, the one for mobility requirements (P two-tail = 0.012) and the one for quality requirements including mobility (P two-tail = 0.042). The other effects were far from being significant.

Table 3 shows the results for the quality of the identified requirements, using the same three-letter variable abbreviations, with ? replacing T or C, that is, QM is the quality of mobility requirements, RX the relevance of other quality requirements (except mobility), CO the clarity of other (nonquality) requirements, and TA the testability over all kinds of requirements. The variables in the grey rows (Quality) are the averages of the ones in the three subsequent white rows (relevance, clarity, and testability). There are only small effects, a few going in favor of the treatment group (the largest one being a 0.2 improvement on the relevance of mobility requirements), others in favor of the control group (the largest, now denoted by a negative number, being a 0.39 advantage for the testability of other quality requirements). On total, it can be seen that it pretty much evens out, the average quality for all kinds of requirements (QA?) has a difference of only −0.03.

An Anderson-Darling test showed that the data for the variables in Table 3 were not normally distributed, hence a Mann-Whitney test was used to investigate significance for some of the largest differences, but none were anywhere near being significant. Hence the null hypotheses cannot be rejected, leading us to conclude that there were no differences in the quality of the requirements between the two groups.

5. Threats to Validity

Wohlin [17] suggests four relevant categories for discussing threats to validity in experiments: conclusion validity, construct validity, internal validity, and external validity. *Conclusion validity* concerns the relationship between the treatment given and the outcome in measured variables. One important question is whether the sample size is big enough to justify the conclusions drawn, which can be investigated by means of the calculated effect size (ES). Our main finding was the acceptance of the hypothesis $H_{1,NM}$, that is, there is a difference in the number of mobility requirements found with the two approaches, for which we found a medium effect with ES = 0.68. We will denote the type I error probability by α and the type II error probability by β. The following relationship holds [21]:

$$N = \frac{4\left(u_{\alpha/2} + u_{\beta}\right)^2}{ES^2}. \quad (1)$$

If we use $\alpha = 0.05$ and $\beta = 0.20$, we get $N = 26/0.68^2$ which gives an N-value of 56. Since we have 54 participants, our sample size is marginally too small, and we should ideally have had a couple more participants in the experiment to make strong claims about the effect.

Construct validity is concerned with the inference from the measures made in the experiment to the theoretical constructs we were trying to observe (understanding, problem solving effectiveness). Our main constructs (variables) were the number and quality of requirements, as well as the participants' preference for the taxonomies offered. Especially for requirements quality, there are many attributes that could be considered, of which we selected only three, namely, relevance, clarity, and testability. This of course does not give a complete picture of requirements quality, but are attributes that the literature has emphasized as important. Also, when it comes to the participants' preference for the

Table 2: Number of requirements found by the participants.

Variable	Treatment group		Control group				
	Mean	St.dev.	Mean	St.dev	Diff	Effect size	Sign.?
NM?	3.39	1.57	2.30	1.46	1.09	0.68	$P < 0.02$
NX?	1.71	1.63	1.46	1.53	0.25	0.16	No
NO?	1.46	1.79	1.77	1.77	−0.31	−0.18	No
NQ?	5.11	2.47	3.77	2.23	1.34	0.55	$P < 0.05$
NA?	6.57	2.44	5.54	2.63	1.03	0.40	No

Table 3: Quality analysis of requirements found by the two group participants.

Variable	Treatment group		Control group				
	Mean	St.dev.	Mean	St.dev.	Diff	Effect size	Sign.?
QM?	1.98	0.73	1.97	0.74	0.01	0.01	No
RM?	2.35	0.82	2.15	0.98	0.20	0.20	No
CM?	1.83	0.99	1.98	0.96	−0.15	−0.16	No
TM?	1.75	0.80	1.77	0.82	−0.02	−0.03	No
QX?	1.08	0.75	1.19	0.72	−0.11	−0.15	No
RX?	0.88	0.96	0.82	0.95	0.06	0.06	No
CX?	1.29	1.03	1.34	0.94	−0.05	−0,05	No
TX?	1.08	0.90	1.42	0.86	−0.34	−0.39	No
QO?	1.17	0.77	1.43	0.79	−0.26	−0.32	No
RO?	0.37	0.58	0.74	1.00	−0.37	−0.37	No
CO?	1.51	1.29	1.80	1.05	−0.29	−0,28	No
TO?	1.63	1.16	1.74	1.08	−0.10	−0.10	No
QA?	1.56	0.85	1.59	0.82	−0.03	−0.04	No
RA?	1.52	1.19	.1.35	1.19	0.17	0.14	No
CA?	1.62	1.09	1.76	1.01	−0.14	−0.14	No
TA?	1.55	0.95	1.67	0.93	−0.12	−0.1	No

taxonomies, there could of course be many measures for this. We based ours on TAM, which has been used a lot in information systems research and validated in a number of studies, this should be a better guarantee for construct validity than inventing some measures all by ourselves. Another threat here is of course that only excerpts of the taxonomies in question were used, this was a necessity due to the limited duration of a controlled experiment.

Internal validity means that the observed outcomes were due to the treatment, not to other factors. Our random distribution experiment design eliminates selection bias and controls for any learning effects or effects of which taxonomy was used. In addition, we also performed a preexperiment questionnaire to test whether other factors such as previous relevant experience could explain the differences between the two groups, not finding any such effects. Another problem could be experimenter bias, for example, the researchers consciously or unconsciously giving a better score to the requirements specifications delivered by treatment group students based on a hope that the treatment would prove successful. This risk was mitigated by having a researcher not involved in any other work with the paper or experiment (Karpati) assess the specified requirements. While Karpati has been included as a coauthor of the paper since his data analysis job took a considerable amount of time, he had no involvement in planning or executing the experiment and had not seen the experimental materials given to the students, only the specified requirements that the students delivered back. Hence he had no knowledge about what treatments were being compared when he did the scoring job, and the student responses were anonymously labeled and contained no information about treatments or treatment groups. The responses were also randomly sequenced with groups intermixed to avoid any bias due to order (e.g., if he had scored all the treatment group responses first, then all the control group, there could have been effects due to fatigue or other similar factors). All in all, we therefore think that internal validity threats have been well addressed.

External validity is concerned with the question of whether it is possible to generalize from the experimental setting to other situations, most importantly to industrial systems development. The use of students instead of practitioners is a notable threat. However, this threat is reduced by the fact that we are only trying to compare two taxonomies in relative terms. Moreover, this mobility taxonomy would be new to practitioners, thus reducing the advantage they might

otherwise have had over students (e.g., if the practitioners had used the taxonomies a lot at work). It should also be noted that [22] found that students have a reasonably good understanding of practice in RE and can therefore be useful experimental subjects, although this would have been even better with master level students. Also, real world requirements engineering is a highly communicative process between the analyst and various stakeholders, while the experiment was individual work transforming a case description to system shall requirements. This threat is not mitigated in the experiment and would have to be solved either by new experiments or by case studies.

6. Conclusion

This paper has reported an experiment comparing some excerpts of the Volere taxonomy with an extension of the same excerpts, the extension especially addressing mobility requirements. The motivation is that such requirements are becoming increasingly important as more and more information systems become mobile and multichannel, the users therefore expecting to be able to perform their tasks in a variety of different locations, with various types of equipment. Taxonomies are believed to be helpful in guiding stakeholders about what requirements might have to be elicited and specified, and the question was therefore whether the extension with mobility requirement would guide the experiment subjects more or better than a taxonomy not including any such category of requirements.

The results showed that the treatment group getting the add-on part about mobility requirements found more such requirements than the control group. This is not surprising or particularly impressive—more guidance should result in better performance. Still, it was not obvious that they would find more requirements, as the extra material might also have been felt as inconsistent with the Volere material or in other ways confusing, causing the participants to lose focus and productivity instead of gaining. Even just the factor of having more tutorial material to read and look through could potentially have resulted in less time spent on effective writing, and thus fewer requirements. So, even if the result is not surprising or impressive, it is still encouraging. However, it must be emphasized that the experiment results by no means prove that "mobility requirement" really deserve to be a category in its own, guidance for writing.

More disappointing, of course, is the finding that the treatment group did not produce better quality requirements than the control group. Ideally, the guidance given in the added tutorial section about mobility should have helped them writing more precise and testable mobility requirements. But quality was not poorer for the control group either, so at least the result for quality has shown that the increased number of requirements found by the treatment group did not come at the cost of reduced quality and so was indeed a productivity increase.

An important question for further work would be to investigate in more detail why the taxonomy only gave an advantage in numbers, not quality. It could be that a quality increase would have been easier to achieve if there had also been tool support for the taxonomy. This could be investigated in a new experiment, but if so, both groups must of course have similar tool support available, only with the difference that the tool of the treatment group offers one extra category of requirements. For new experiments, it would also be interesting to have more complex cases and maybe have participants working in pairs, for example, one domain expert and one analyst, so that the experimental task becomes more similar to a realistic working situation.

Finally, it would also be important to perform larger case studies in mobile information systems projects, preferably in industry, to check whether advantages observed in a limited experimental setting also hold for real world usage. Such industrial evaluations would also give more insight on the workplace usefulness of applying taxonomies for eliciting requirements.

References

[1] J. Robertson and S. Robertson, *Volere Requirements Specification Template*, Atlantic Systems Guild, 2010.

[2] D. G. Firesmith, *Common Concepts Underlying Safety, Security, and Survivability Engineering*, Carnegie Mellon University, Software Engineering Institute, Pittsburgh, Pa, USA , 2003.

[3] D.G. Firesmith, "A taxonomy of safety-related requirements," in *Proceedings of the Workshop on Requirements for High Assurance Systems, (RHAS '04)*, p. 11, IEEE Computer Society, Kyoto, Japan, September 2004.

[4] D. G. A Firesmith, "Taxonomy of security-related requirements," in *Proceedings of the International Workshop on High Assurance Systems, (RHAS '05)*, Paris, France, August 2005.

[5] A. I. Antón, J. B. Earp, and A. Reese, "Analyzing website privacy requirements using a privacy goal taxonomy," in *Proceedings of the IEEE Joint International Requirements Engineering Conference*, K. Pohl, Ed., pp. 23–31, IEEE, Essen, Germany, 2002.

[6] M. Mahaux, P. Heymans, and G. Saval, "Discovering sustainability requirements: an experience report, in requirements engineering," in *Proceedings of the 17th International Working Conference, (REFSQ '11)*, D. Berry and X. Franch, Eds., Lecture Notes in Computer Science, pp. 19–33, Springer, Essen, Germany, March 2011.

[7] G. Sindre, "Trust-related requirements: a taxonomy," in *Advances in Information Systems Development*, G. Magyar, Ed., pp. 49–61, Springer, Berlin, Germany, 2007.

[8] R. MacRuairi, M.T. Keane, and G. Coleman, "A wireless sensor network application requirements taxonomy," in *Proceedings of the 2nd International Conference, (SENSORCOMM '08)*, Sensor Technologies and Applications, pp. 209–216, Cap Esterel, August 2008.

[9] S. Gopalakrishnan and G. Sindre, "Taxonomy of mobility-related requirements," in *Proceedings of the International Conference on Interoperability for Enterprise Software and Applications, (IESA '09)*, pp. 283–288, Beijing, China, April 2009.

[10] S. Gopalakrishnan and G. Sindre, "A revised taxonomy of mobility-related requirements," in *Proceedings of the International Conference on Ultra Modern Telecommunications and Workshops, (ICUMT '09)*, pp. 1–7, St. Petersburg, Russia, October 2009.

[11] S. Gopalakrishnan and G. Sindre, "A study on mobile requirements elicitation by boilerplate requirements specification

language," in *Proceedings of the Proceedings of the International Conference on Electronic Business*, pp. 613–623, 2010.

[12] E. Hull, K. Jackson, and J. Dick, *Requirements Engineering*, Springer, London, UK, 2010.

[13] J. K. Gershenson and L. A. Stauffer, "Assessing the usefulness of a taxonomy of design requirements for manufacturing," *Concurrent Engineering Research and Applications*, vol. 7, no. 2, pp. 147–158, 1999.

[14] K. S. Rounds and J. S. Cooper, "Development of product design requirements using taxonomies of environmental issues," *Research in Engineering Design*, vol. 13, no. 2, pp. 94–108, 2002.

[15] D. T. Haley, B. Nuseibeh, H. C. Sharp, and J. Taylor, "The conundrum of categorising requirements: managing requirements for learning on the move," in *Proceedings of 12th IEEE International Requirements Engineering Conference, (RE '04)*, pp. 309–314, September 2004.

[16] M. Glinz, "On non-functional requirements," in *Proceedings of the 15th IEEE International of Requirements Engineering Conference, (RE '07)*, pp. 21–26, October 2007.

[17] C. Wohlin, *Experimentation in Software Engineering: An Introduction*, Kluwer Academic, Norwell, Mass, USA, 2000.

[18] E. Knauss and C. El Boustani, "Assessing the quality of software requirements specifications," in *Proceedings of the 16th IEEE International Requirements Engineering Conference, (RE '08)*, pp. 341–342, Catalunya, Spain, September 2008.

[19] D. Firesmith, "Specifying good requirements," *Journal of Object Technology*, vol. 2, no. 4, pp. 77–87, 2003.

[20] K. Winbladh, H. Ziv, and D. J. Richardson, "Eliciting required characteristics for usable requirements engineering approaches," in *24th Annual ACM Symposium on Applied Computing, (SAC '09)*, pp. 360–364, Honolulu, Hawaii, USA, March 2009.

[21] W. G. Hopkins, *A New View of Statistics*, University of Queensland, Brisbane, Australia, 2001.

[22] M. Svahnberg, A. Aurum, and C. Wohlin, "Using students as subjects—an empirical evaluation," in *Proceedings of the ACM-IEEE international symposium on Empirical Software Engineering and Measurement*, pp. 288–290, ACM, Kaiserslautern, Germany, 2008.