Confounding Factors When Conducting Industrial Replications in Requirements Engineering

David Callele
Business Development
Telecommunications Research Laboratories (TRLabs)
Saskatoon, Canada
dcallele@trlabs.ca

Krzysztof Wnuk
Department of Computer Science
Lund University
Lund, Sweden
Krzysztof.Wnuk@cs.lth.se

Markus Borg
Department of Computer Science
Lund University
Lund, Sweden
Markus.Borg@cs.lth.se

***Abstract*— Despite the widely recognized importance of replications in software engineering, industrial replications in software engineering are still rarely reported. Although the literature provides some evidence about the issues and challenges related to conducting experiments and replications the practitioner's view of the issues and challenges has not been fully explored. This paper reports an industrial practitioner's review of a replicated experiment on linguistic tool support for consolidation of requirements from multiple sources. The review identified potential confounding factors from a perspective that differed significantly from that of the designers of the experiment. The results suggest that industrial practice may focus upon specific process aspects that are not necessarily reflected in academic practice.**

Keywords: Requirements engineering, replication, confounding factors, experience report.

I. INTRODUCTION

Replications play an important role in software engineering, furthering our knowledge about which results or observations hold and under what conditions [1]. Despite the widely recognized importance of this type of research study, replications in software engineering are still rarely reported; Sjøberg *et al.* [2] reported that replication studies constituted only 18% of the surveyed experiments and that only 9% of the subjects in the reviewed experiments were practitioners. Furthermore, undergraduate students are used much more often than graduate students [2], while conducting an industrial replication could bring invaluable insights whether the observations hold under real industrial conditions. Moreover, the comprehensive software engineering experimentation literature that identifies threats to validity analysis (*e.g.* [11]) does not discuss whether the reported threats are valid for industrial replications as well as academic replications.

In this paper we report our experiences when replicating an experiment in automatic support for finding and recording similar requirements that originate from different customers [3]. The automatic support tool used the cosine correlation measure to find and present lexically similar requirements to the human analyst [3]. The experiment used students as subjects and investigated whether a tool with linguistic similarity functionality can help the subjects to analyze more requirements, identify and create a greater number of correct requirements links, and miss fewer links than a tool without this automated support.

Our main research question for this study is:

"*Are there additional confounding factors that should be taken into consideration when replicating an experiment in industry?*"

An industrial practitioner review of a replicated experiment in automated support for detecting similar requirements that originate from different customers [3][4] identified confounding factors relevant to an industrial replication of the experiment.

This paper is structured as follows: Section II outlines the context of the original and replicated experiments; Section III reports on the results from both experiments and how their discussion triggered the industrial practitioner's review of the replicated experiment. Section IV reports on findings from the analysis of the work done by an independent industrial project manager interested in the possible benefits of applying the experimental results in practice and Section V concludes the paper.

II. REPLICATION CONTEXT

This study assumes that incoming requirements and changes to existing requirements are inevitable throughout the entire development process [5] and that requirements originate from multiple sources [3][4][6][7]. Having many customers (sources of requirements) creates a risk that some incoming requirements will be similar to already implemented requirements or considered, but not yet implemented, requirements. For large companies, operating globally, the list of requirements sources is long and the stream of incoming requirements can easily overwhelm the capacity of the requirements analysts receiving them. The requirements analysts need to check every incoming requirement to identify whether or not it was already investigated or implemented. One possible way to assist in this analysis is to find and record similarities while making traceability links. Finding similar requirements has been reported as a time consuming and frustrating task [4].

Requirements sources (customers and subcontractors) do not necessarily have knowledge of what requirements may already have been received or knowledge of what requirements have already been implemented. The goal is to

rationalize these incoming requirements, identifying their sources, their similarities, whether they have been analyzed already (possibly in a similar but not identical form) and whether they have already been implemented. As the number of requirement sources grows, and the number of received requirements grows, the task of linking them becomes exponentially more challenging.

The conceptual solution to this activity is presented in Figure 1 (also available in [3][4]). To the left in Figure 1 two versions of a requirements specification from the same key customer are shown (A and B). Sets A and B could represent requirements from different customers or A could be an earlier specification and B is a later version of the specification. The task is to analyze the requirements in the B set against the requirements in the A set. Requirements analysts could manually perform this time-consuming analysis but the goal of both experiments was to introduce automated support for this task.

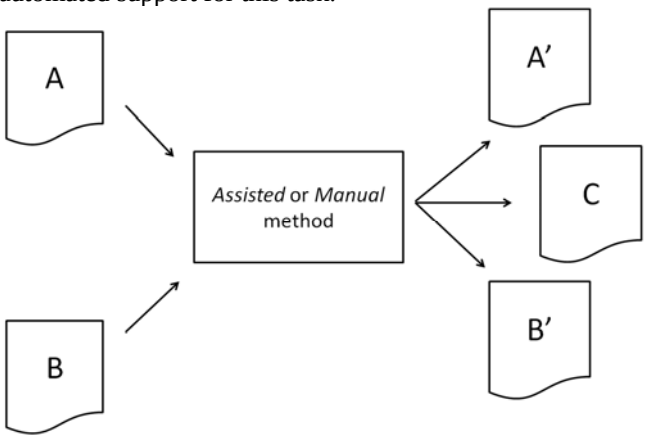

Figure 1. The conceptual solution to the activity of finding similar requirements (also available at [3][4]).

Subjects in both experiments used two supporting methods: automated (ReqSimile in both cases) or manual (ReqSimileM or Doors) to find requirements in set B that were already analyzed in set A and to mark them by assigning a link between them. The output of the process comprises subsets: A' (all requirements that are not linked to any requirement in the set A), B' (all new requirements that have not previously been analyzed) and C (all requirements in the new specification that previously have been analyzed). The B' set is sent to further analysis and the return for using the method is the time saved by not analyzing subsets A' and C.

An automatic method for analyzing similarity between incoming requirements could significantly decrease the amount of time needed to perform this task. Both the original experiment, reported in [4] and the replicated study analyzed in this paper [3] attempt to assess the benefits of using a linguistic similarity method against a manual method for finding and linking similar requirements.

The linguistic similarity method used in both experiments (ReqSimile [9]) measures lexical similarity between requirements, ranking candidate requirements for linking then presenting to the user with the most relevant of those requirements. ReqSimile utilizes the cosine correlation measure, where each requirement is represented by a vector of linguistic terms with the respective number of occurrences of each term [6][8].

The second method differed between the two experiments. In the original experiment [4] the second method, also called the *manual method*, was represented by a modified version of the ReqSimile tool [7][9] where the linguistic similarity functionality was disabled and participants were constrained to using simple keyword searching to identify similarities. In the replicated experiment [3], the manual method was represented by searching and filtering functionality provided by the Telelogic DOORS tool [10]. In addition to keyword searching, DOORS allows the user to select the attributes included in the search, or to use UNIX-style regular expressions. Search results may also be filtered by attribute and content and simple filters may be combined to create more complex filter operations.

The remaining variables in this experiment were kept as close to unchanged from the original study as could be managed and can be accessed in the original and replicated experiment publications [3][4]. The remaining principal difference between the original and the replication was that participants worked in pairs during the replicated experiment rather than singly as in the original.

As noted above, the independent variable is the method used by participants in the experiment. The controlled variable is the experience of the participants, evaluated by a questionnaire. The dependent variables considered are:

(1) time used for the consolidation,
(2) number of analyzed requirements,
(3) number of correct links,
(4) number of incorrect links,
(5) number of correctly not linked and
(6) number of missed links (incorrectly not linked).

The number of analyzed requirements (2) is used in case the subjects are not able to analyze all requirements, which will affect the results of (5) and (6). The hypotheses for comparing the manual and the assisted method remain unchanged from the original experiment [4]:

(H_1)The assisted method results in the same number of requirements analyzed per minute, N/T, as the manual method.

(H_2) The assisted method results in the same share of correctly linked requirements as the manual method.

(H_3) The assisted method results in the same share of missed requirements links as the manual method.

(H_4) The assisted method results in the same share of incorrectly linked requirements as the manual method.

(H_5) The assisted method is as precise as the manual method.

(H_6) The assisted method is as accurate as the manual method.

The replicated experiment used a different set of participants, drawn from a similar population – a course in Requirements Engineering. The course is an optional master-level course offered for students in several programs including computer science and electrical engineering. Although the experiment was a mandatory part of the course, the results achieved by the subjects had no influence on their final grade from the course. There were 45 subjects participating in the replicated experiment and 44 participating in the original experiment. Before conducting the experiment, the subjects had been taught requirements engineering terminology and had gained practical experience through their course project. Two questionnaires were used, before and after conducting the experiment to record the subjects' skills in reading and writing English and their industrial experience in software development. The participants in both experiments used two requirements sets, one written in case style (139 requirements) and the other in feature style (160 requirements). The participants were asked to analyze 30 randomly selected requirements from the first set against all 160 requirements from the second set, and to create links where necessary.

III. Resulting from Both Experiments Triggering Industrial Practitioner's Review

The results from hypotheses testing in both experiments are summarized in Table I. 'Significant' means that there was a statistically significant difference (with a 95% confidence level) between the manual and the assisted method. For non-significant results, we put interpretations from the replicated experiment as the original experiment reported few interpretations [4]. In the case of H_1 the results from the replicated experiment were not consistent with the original experiment. The possible interpretations outlined in [3] included:

(1) the large variation of the results for the *assisted* method in the replicated experiment (minimum value of 0.29 and maximum value of 0.89 requirements per minute),

(2) the fact that subjects experienced in reviewing requirements were not the fastest (achieving scores near the median) and

(3) the fact that the two lowest values for performance were achieved by the subject pair with industrial experience.

The results for H_2 and H_3 (significant in both original and replicated experiments) could be interpreted in favor of the assisted method (ReqSimile) as it performed significantly better even against a rather sophisticated requirements management tool (Doors) 0 in regard to correctness and missed requirements links.

Both original and replicated experiments did not provide statistically significant results regarding the number of incorrect links (H_4), precision (H_5) and accuracy (H_6). In the replicated experiment these were interpreted as a sign of additional (unidentified) confounding factors that affect the consolidation of requirements and were not controlled in the experiment. The lack of statistical significance regarding H5 and H6 was explained in the original experiment as caused by the fact that both precision and accuracy incorporated the number of incorrectly linked requirements. As the number of incorrect requirements turned out to be large in relation to the other measures, the authors of the original experiment suggested that this may be the reason of lack of statistical significance [4].

Table 1. The result from hypotheses testing in both experiments. '!significant' indicates that the results from hypothesis testing were *not* significant (with a 95% confidence level).

Hypothesis	Result original experiment	Result replicated study	Possible interpretation (see also [3])
H_1: Speed	significant	!significant	Observed wide variation in results, possibly due to participant motivation
H_2: Correctness	significant	significant	
H_3:Missed links	significant	significant	
H_4:Incorrect links	!significant	!significant	Uncovered confounding factors such as pairs *vs.* singleton and (minimal) industrial experience.
H_5: Precision	!significant	!significant	
H_6: Accuracy	!significant	!significant	

Triggered by the replicated lack of statistical significance regarding hypotheses H_4, H_5, H_6, and possible additional factors affecting the results regarding H_1, we performed an independent review by an industrial practitioner who reviewed the experimental design and results from both experimental runs. The goal of the review was twofold:

(1) to search for additional confounding factors that may be important when replicating the experiment in industry

(2) to seek further explanations for the lack of statistically significant results regarding H_4, H_5 and H_6. The results from this review are outlines in the section that follows.

IV. Result Review by Industrial Practitioner

The first author reviewed the design of, and the results from, the experiment (after its publication) to determine whether he could identify additional confounding factors that should be considered when replicating the experiment in industry, identifying the following activities and decisions as potential confounding factors.

A. Task Analysis

Participants in the study were required to perform a series of analysis tasks [2] with aspects that are mechanical (such as reading through lists, selecting from lists and performing sequences of operations to create logical links between list elements) and aspects that are cognitive (such

as interpreting statements, remembering statements and correlating between sets of statements).

We present three subtasks that have the potential to be significant confounding factors when interpreting the results of the experiment: the complexity of generating the search terms, the number of search results returned by a query and the complexity of interpreting the search results. These subtasks are illustrative and not necessarily exhaustive.

1) Complexity of generating the search terms

There is a significant difference in the number of operations necessary to generate the search results between the two evaluated methods. In the assisted method, the work is performed by the underlying tool [7][9]. However, in the manual case, the requirement in question must be analyzed by the participants and appropriate search terms must be generated. Records of the search terms were not kept, nor were the number of attempts made by the participants to generate the final working set recorded. Even if the participants in the manual study were able to generate their final working set on their first attempt, the complexity of the task is much higher than the assisted method. Given the time constraints for the experiments, does the effort for manual generation of the search terms overwhelm the other results? We do not see direct evidence to confirm or deny that this confounding factor actually occurred in these experiments. This confounding factor is related to the inadequate preoperational explication of construct threat [11] but is not discussed in the recent literature study [2]. This confounding factor could influence the results regarding incorrect links (H_4) precision (H_5) and accuracy (H_6).

2) Number of search results in relation to the 'quality' of the search terms.

Each approach attempts to constrain the search for requirements links from an n*n search space (for n requirements) to a space m*n where m is the number of search results and m is expected to be less than, or much less than, n. The smaller the resulting search space m (without missing necessary candidates), the more efficient the work process can be.

We are not able to determine the 'quality' of the search terms employed by each team and the number of search results presented to each participant pair could be significantly different. As a result, any differences in the measured results may have been caused by a significant imbalance between the numbers of requirements that were presented to the participants. There is a risk that the experiment is measuring each team's ability to generate effective and efficient search terms and this factor may dominate other results. This factor is related to the reliability of used measures threat [11] and might have been one of the factors affecting the number of incorrect links (H_4), precision (H_5) and accuracy (H_6).

3) Complexity of interpreting search results

We assume that the assisted method returned the same result set to each team. Therefore, differences in results for each team may be attributed to other factors. For example, the order in which participants addressed the requirements can have a significant impact upon the results. However, we do not have the same degree of control over the manual method. Furthermore, the fact that a list of highly similar requirements sorted by their similarity degree was presented to subjects may potentially increase the number of incorrect links (H_4) as more false positives are generated, which in turn may negatively impact precision (H_5) and accuracy (H_6).

Let us assume that the requirements are of varying levels of complexity and with varying levels of linkage to other requirements. Given this assumption, and under the knowledge that the experiment is time-constrained, then the order in which participants addressed the requirements can have a significant impact upon the results.

The reported time for running the experiments was 45 minutes, or 90 seconds per requirement – far less than in industrial practice. To facilitate industrial adoption of the results, practitioner time-constraints should have been removed or kept to a reasonable approximation of the effort deemed acceptable to industrial practice. This factor conflicts with the history and maturation threats to internal validity [11]. This confounding factor could influence the results regarding performance (H_1).

B. Further confounding factors

1) Using students as practitioner proxies

This factor has been listed among the threats to internal and external validity by Wohlin et al. [11] and discussed by several researchers e.g. [12]. In the first author's practice, none of the replication subjects would likely be considered to have sufficient experience to participate in a requirements effort, except as support staff "in training", which impacts the credibility of the results. The most experienced participants claimed approximately two years of experience, none of which included a focus on requirements. It was not identified whether that experience was two consecutive years or two cumulative years. Interestingly, the analysis of the possible influence of industrial experience on results revealed that, in most cases, this factor negatively affected results [3].

To place this issue in context, the practice of engineering in Canada is regulated by statute in each of the 10 provinces. While there are subtle differences across the provinces, engineers are expected to have four years of experience as Engineers-In-Training, under the guidance of a senior engineer with Professional Engineering accreditation, before they can apply for Professional Engineer status themselves. This experience must include the application of theory and should provide exposure to, or experience in, the following broad areas: practical experience, management, communication, and the social implications of engineering [13]. Requirements engineering is an engineering management task; novice engineers-in-

training may be introduced to management tasks but are unlikely to be assigned responsibility for these tasks until late in their training. In the current case, it is the opinion of the first author that the claims made for the results of this experiment are valid for practitioners with little or no experience but generalizing to experienced practitioners is difficult to justify.

2) Reading speed

The reading speed of the participants can be a significant factor in a timed evaluation [14] and could influence performance results (H_1). Adult reading speeds, with significant comprehension, are widely reported from about 100 words per minute to approximately 1000 words per minute, with scanning speeds even higher. A practitioner with high-speed scanning skills would be expected to perform elements of the task at rates of up to an order of magnitude more quickly than their slowest counterparts. If the experimental results do not compensate for this aspect of the environment, they can be dominated by this one factor alone. This factor isn't explicitly mentioned by Wohlin *et al.* [11]. This confounding factor could influence the results regarding correctness (H_2) and accuracy (H_6).

3) Solution strategies undertaken by subjects

The solution strategy undertaken by each team may not be a linear scan through the presented alternatives. Strategies that may have been employed by teams to improve their performance over a linear scan include, but are not limited to:

- Partition the requirements between the partners, with each partner working independently. This approach is particularly effective for generating the search terms in the manual approach.
- Partition the result sets between the partners, one partner working from the bottom of the result set toward the top, the other from the top down.
- Partition the result sets by length of requirements. Scan and evaluate all short requirements first, then proceed to the longer requirements.
- Process all requirements stated using a simple sentence structure first, then move on to compound and complex sentence structure requirements.

These strategies could influence the results regarding performance (H_1) precision (H_5) and accuracy (H_6). If the participants employed any of these techniques (or others not explicitly mentioned here) and did not accurately report upon the technique used in the post-experiment questionnaire then the results may be biased. This factor isn't explicitly mentioned by Wohlin *et al.* [11].

4) Personalities

Against intuition, the replicated experiment [3] did not find that experienced participants outperformed inexperienced participants. Personality may have played a significant role in the experiment. Pairs of dominant personalities may have clashed, pairs of passive personalities may have dithered and mixing a dominant and a passive personality may have been a waste of the passive participant resource. This factor was not considered in the analyzed experiment and is often not considered in industrial environments despite the significant potential for the noted issues to occur.

The first author posits that the inexperienced participants may also have been afraid of failure or appearing incompetent in front of their peers. As a result, their attentiveness was boosted, as was their attention to detail. The relatively experienced participants may have been preconditioned by their (meager) experience and they may have been overconfident. We must remember that the study participants are actually novices, freshmen in second term, not even second year sophomores. This factor isn't explicitly mentioned by Wohlin *et al.* [11].

5) Sampling validity

The population of 160 requirements was sampled to create a working set of 30 requirements. The contents of this working set may or may not have been representative of the greater population. The requirements themselves were not analyzed to determine what effect changing the working set would have upon the experimental results. Confounding factors include the complexity of the requirement statements, the nature of the requirements (*e.g.* functional *vs.* non-functional) and the requirements domain (well-understood or not). Wohlin *et al.* focuses primarily on strategies for subject selection [11].

V. CONCLUSIONS

This paper has reflected upon some challenges associated with performing replications of empirical software research experiments in an industrial setting. The analysis indicates that industrial practice may focus upon specific aspects of processes that are not necessarily reflected in academic practice. For example, human factors such as reading speed and personalities, or process optimizations such as solution strategies may be ignored when designing academic experiments as they are not explicitly listed in the experimentation literature [2][11].

However, these factors may need to be addressed to justify applying the research results in an industrial environment or to obtain investment in equivalent industrial experiments. We note that our analysis is based on the opinion of a single practitioner and there exists a risk that our findings are specific to the type of experiment that was conducted. However, we believe that industrial practice is necessarily holistic and that our experiments should consider sensitivity to environmental aspects. Thus, we encourage researchers to perform a sensitivity analysis to aspects such as those detailed here.

REFERENCES

[1] F. J. Shull, S. Vegas, N. Juristo, "The role of replications in empirical software engineering," Empirical Software Engineering, vol. 13, Apr. 2008, pp. 211–218, doi: 10.1007/s10664-008-9060-1.

[2] D. I. K. Sjøberg, J. E. Hannay, O. Hansen, V. B. Karahasanovic, A. Liborg, N. K. Rekdal, "The survey of

controlled experiments in software engineering," IEEE Trans Softw Eng, vol 31, Sep. 2005, pp. 733–753, doi: 10.1109/TSE.2005.97.

[3] K. Wnuk, M. Höst, B. Regnell, "Replication of an Experiment on Linguistic Tool Support for Consolidation of Requirements from Multiple Sources," Empirical Software Engineering, 17(3) pp. 305-344, June 2012, doi: 10.1007/s10664-011-9174-8.

[4] J. Natt och Dag, T. Thelin, B. Regnell, "An experiment on linguistic tool support for consolidation of requirements from multiple sources in market-driven product development," Empirical Software Engineering, vol. 11, Jun. 2006, pp. 303–329, doi:10.1007/s10664-006-6405-5.

[5] G. Kotonya and I. Sommerville "Requirements Engineering: Processes and Techniques" , Wiley, 1998.

[6] J. Natt och Dag, V. Gervasi, S. Brinkkemper, B. Regnell, "Speeding up requirements management in a product software company: Linking customer wishes to product requirements through linguistic engineering," Proc. 12th International Requirements Engineering Conference (RE 2004), IEEE Press, Sept. 2004, pp 283–294, doi: 10.1109/ICRE.2004.1335685.

[7] J. Natt och Dag, Managing natural language requirements in large-scale software development. PhD thesis, Lund University, Sweden, 2006.

[8] C. D. Manning, H. Schuetze, Foundations of Statistical Natural Language Processing. MIT Press, 2002.

[9] The Reqsimile project is available at http://reqsimile.sourceforge.net/

[10] IBM Rational doors product description (former Telelogic Doors), accessed March 2011. http://www-01.ibm.com/software/awdtools/doors/productline/

[11] C. Wohlin, P. Runeson, M. Höst, M. Ohlsson, B. Regnell, A. Wesslén, "Experimentation in Software Engineering," Springer, 2012.

[12] M. Höst, B. Regnell, C. Wohlin, "Using Students as Subjects – A Comparative Study of Students and Professionals in Lead-Time Impact Assessment," Emp Soft Eng, vol. 5, pp. 201-214, Nov. 2000.

[13] APEGS, Association of Professional Engineers and Geoscientists of Saskatchewan Components of Acceptable Engineering Work Experience (Engineers Canada Interpretive Guide IV). Experience Guideline 2.

[14] Hudson, R.F., Lane, H.B., & Pullen, P.C. (2005). Reading fluency assessment and instruction: What, why, and how? The Reading Teacher. 58, No.8.