Transitioning from lab studies to large-scale studies: Emerging results from a literal replication

Remo Ferrari, Nazim H. Madhavji
The University of Western Ontario, Canada
{rnferrar, madhavji}@csd.uwo.ca

Oliver Sudmann, Christian Henke, Jens Geisler, Wilhelm Schaefer
University of Paderborn, Germany
{oliversu, chenke, jgeisler, wilhelm}@upb.de

ABSTRACT
Replication of studies in Software Engineering is considered important, but is largely neglected. Because of the lack of published replicated literature, there are few established guidelines for researchers wanting to conduct replicated studies. Specifically, guidelines for transitioning from laboratory studies to large-scale studies are nonexistent. Previously, we conducted a laboratory study in the banking domain, in Canada, which we replicated by conducting an extended, large-scale case study on an innovative rail project in Germany, investigating the role of an existing systems architecture on requirements decisions. In this short paper, we present a preliminary analysis of our transitioning experiences from conducting these two studies. From our experiences, we derive a set of lessons learnt and recommendations that can be used by other researchers wanting to transition from lab studies to studies in industrial settings.

1. INTRODUCTION
Though replication plays an important role in scientific studies, there is a severe lack of published replicated studies in Software Engineering (SE). In a recent survey published by Sjoberg et al. [6] only 20 instances of published SE replication studies exist. To date, the few other researchers who have attempted replication in SE (see, e.g., [6]) have conducted “literal” replication (i.e., replicated study which aims to produce results that are directly comparable to the original study, though the study may differ in several dimensions such as context, environment, participants, etc. [8]).

In this paper, we examine a complementary issue – that of “transitioning” from lab studies to large-scale studies, specifically when using the case study methodology. Our observations are rooted in two studies.

The first is the, original, lab study [5], that was conducted in an academic setting in Canada using university students as participants, observing a two-month long Requirements Engineering (RE) project in the banking domain. The second is the replicated case study [3] that was much larger in scale and collected data from an experimental Rail system being developed in Germany for more than 10 years, with a budget in excess of 11 million Euros. This project employs full-time practitioners, along with students and professors.

The replicated case study overlaps with the original lab study in that it contains the key investigative question: “**In which manner does a systems architecture (SA) affect requirements decisions?**” However, there are two major differences as well: (i) scale (as described above) and (ii) extension of the investigation. The extension culminates into two new investigative issues: (a) frequency and type of the source of the requirements decision (e.g., external stakeholder request, feedback from implementation activities, emergent, etc.) and (b) the impact of the affected requirements decisions on downstream implementation activities. These two issues could not be investigated in the original lab study.

The motivation for such transitioning analysis is rooted in the fact that many lab studies never make it to industrial-scale practice because they are still preliminary. Simply put, you do not fly passengers in a prototype plane! Our transitioning analysis can thus enable or encourage other researchers to transition lab studies to large-scale case studies. The end objective is that, the more lab studies are transitioned to large-scale case studies, the more SE practice is likely to use empirical results in driving software development and maintenance [4].

To the best of our knowledge, there is no published paper that focuses on such transitioning. The output of this short paper is a preliminary set of lessons learnt and recommendations that could be used in transitioning efforts.

The paper is structured as follows: in Section 2 we discuss related work on replication in general; in Section 3 we describe our transitioning experiences; in Section 4 we describe lessons learnt and recommendations based on our experiences, and Section 5 concludes the paper and discusses future work.

2. Related Work
In this subsection, we describe some background information on empirical study replication. Because the replication conducted in software engineering (SE) has been sparse, this subsection focuses on general empirical literature outside of SE.

Yin, in [8], defines three different types of replication that can be conducted: exact, literal and theoretical. Exact replication is a direct copy of the original study, including constructs, measurements, environment, etc. However, Brookes et al. in [1], suggest that exact replications involving human participants in SE is not possible, because at least some variations (e.g., participants background, working environment, study duration, etc.) will exist, which will lead to studies that are in fact not exact duplicates, but “literal” replication studies.

Thus, the more common type of replication is literal, where the same constructs and basic measurements are used but different cases are selected that, based on the similarity of the cases used, the results are directly comparable to the original study [8]. The results do not necessarily need to be the same but any differences should be explainable based on differences in the properties of the case selected. Thus, in these types of studies, it is critical that the researcher notes the similarities and differences between the original study and replicated study cases, and argue the appropriateness of the new cases selected for a literal replication. In the large-scale replicated study discussed in this paper, it is categorized as a literal replication.

The last type of replication is a “theoretical” replication, where the replicated study predicts contrasting results to the original but for anticipatable reasons, usually operationalized through a change...
in the research questions or constructs [8]. The theoretical replication is complementary to the literal replication and is used to more fully describe the theoretical framework of the phenomena being investigated, to state more precisely under which specific conditions the phenomena will occur. Typically a theoretical replication will occur after several instances of literal replication have been conducted. Note that the theoretical replication type of study does not necessarily encompass extensions to the original study (such as additional, orthogonal research questions as used in our study) - these are simply extended literal studies.

3. Transitioning Experiences
In this section, we describe our preliminary experiences when transitioning from the lab-study to the large-scale case study. In the next subsection, we discuss the case study attributes that we compare and discuss. Following this, we provide a general description of the two studies, and lastly, we describe the specific transitioning experiences.

3.1 Transitioning attributes
Table 1 lists the key attributes that are used to compare and contrast between the lab study and the large-scale study, and their data values. They are clustered around the themes: case study context and case study design. The attributes contained within the “case study context” theme (e.g., project size, budget, domain, etc.) are typical software project, product and process attributes [7]. The attributes in the “case study design” theme are all recognized case study procedural “steps” (such as data collection or participant selection) and “issues and elements” (such as threats to validity and data sources) [8]. Another theme that is not described in this short paper include such items as degree of researcher control, ethics and accountability. Note that it is quite possible that for other empirical studies, there may be yet more attributes of interest.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Study characteristics</th>
<th>Lab study</th>
<th>Large-scale case study (RailCab)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case study context</td>
<td>Project Scale</td>
<td>Class project</td>
<td>Demonstrable prototype system</td>
</tr>
<tr>
<td></td>
<td>Developer-Type</td>
<td>University students</td>
<td>Mix of professionals, professors, and students (graduate)</td>
</tr>
<tr>
<td></td>
<td>Project budget</td>
<td>CAD $&lt;10K or &lt;8K Euros (April, 2010)</td>
<td>Approximately 11 million Euros</td>
</tr>
<tr>
<td></td>
<td>Development cycle time</td>
<td>2 month project</td>
<td>Over 10 years</td>
</tr>
<tr>
<td></td>
<td>Team Size</td>
<td>2 undergraduate students</td>
<td>Approximately, a mix of 50 full and part-time professionals, professors, students and researchers.</td>
</tr>
<tr>
<td></td>
<td># of cases</td>
<td>6 architecting teams conducting the same project</td>
<td>1 project, with investigation conducted on 3 major system modules.</td>
</tr>
<tr>
<td></td>
<td>Application Type</td>
<td>Banking</td>
<td>Experimental Rail System</td>
</tr>
<tr>
<td></td>
<td>Study Start</td>
<td>Start of project</td>
<td>Midst of project</td>
</tr>
<tr>
<td></td>
<td>Language of project documentation</td>
<td>English</td>
<td>Predominantly German, some English</td>
</tr>
<tr>
<td></td>
<td>Project staff turnover</td>
<td>Not existent</td>
<td>Moderate level of turnover (especially with graduate students)</td>
</tr>
<tr>
<td></td>
<td>Source of project</td>
<td>Class instructor.</td>
<td>Govt. funding, external organizations, University initiative.</td>
</tr>
<tr>
<td>Study design</td>
<td>Investigation questions</td>
<td>Q1 (see section 3.2.1)</td>
<td>Q1-3 (see section 3.2.2) (Q2 -3 not possible in lab study)</td>
</tr>
<tr>
<td></td>
<td>Constructs</td>
<td>Requirements decision, RE-SA interaction</td>
<td>(a) Requirements decision, RE-SA interaction, (b) impact on downstream activities, and (c) source of decision.</td>
</tr>
<tr>
<td></td>
<td>Participant selection</td>
<td>students (availability sampling)</td>
<td>Professionals chosen for their expertise and impact.</td>
</tr>
<tr>
<td></td>
<td>Data Sources</td>
<td>Student developers, data logging tool, project documents.</td>
<td>Project documents, developers, running system and development environment, emails, phone calls.</td>
</tr>
<tr>
<td></td>
<td>Data collection procedures</td>
<td>Students capture data in the logging tool; meetings with individual groups; submitted project documentation</td>
<td>Identifying key project documents and obtaining copies of these; German to English translation and verification through triangulation; audio recordings of semi-structured interviews conducted; email exchanges; phone calls.</td>
</tr>
<tr>
<td></td>
<td>Analysis procedures</td>
<td>Simple case study descriptive analysis (categorization of decisions and frequency counts).</td>
<td>Qualitative analysis (content analysis) of project documents and audio transcripts. Simple case study descriptive analysis (categorization of decisions and frequency counts).</td>
</tr>
<tr>
<td></td>
<td>Threats to validity</td>
<td>Using students as participants, project not implemented</td>
<td>Retrospective analysis, specialized domain.</td>
</tr>
</tbody>
</table>
3.2 Project descriptions

3.2.1 Lab study

The lab study project [5] was conducted as part of a Requirements Engineering course at the University of Western Ontario. The main research question (Q1) investigated how new requirements decisions during RE were affected by the presence of an existing SA. The participants of the study were final-year undergraduate students. The study involved six RE teams that were observed over the course of two months. The participants were given a set of tasks that involved upgrading the requirements for an existing banking system class project. For this purpose, they were given the pre-existing requirements and architecture documents from the previous version of the system. To collect data, we developed a software tool. The study’s findings [5] were that there were four different ways in which an SA affected requirements decisions and their quantitative profile: (i) constrained (25%): where the SA made an RE decision less- or infeasible; (ii) enabled (30%): where the SA made an RE decision (more) feasible; (iii) influenced (6%): where the SA changed an RE decision without affecting its feasibility; and (iv) none (39%): where the SA had no effect.

3.2.2 Large-scale study

The large-scale case study was conducted as a literal replication [8] of the lab study, meaning that not only was the basic investigation of [5] (the impact of an architecture on requirements decisions) conducted but, also, it extended the lab study by investigating: (Q2) the source and type of the requirements decisions, and (Q3) the impact of the affected decisions on downstream development activities and the resultant system. The case investigated for this replication study was the RailCab project [3]: an experimental Rail system that has been in development for approximately ten years at the University of Paderborn in Germany, with a budget of approximately 11 million Euros.

The goal of the project is to create a comprehensive concept for a future railway system. The key feature of RailCab is that it is an autonomous, self-adapting system, and thus does not require any human operator to drive the train. The train’s test track is approximately 530 metres in length with one track-switch and one railway station, and the RailCab vehicles are constructed to the scale of 1:2.5. The RailCab consists of five major components: Drive and Brake, Energy Management, Active Guidance and Steering, Tilt and Suspension, and Motor and Track Design.

The case study involved examining project documents and extensively interviewing eight RailCab staff members regarding the RE-SA interactions; this process was executed over the span of approximately one-year.

The results [3] showed that 25% of the decisions were constrained, 12% were enabled, and 65% were not affected. Further to this, the constrained decisions were observed to have a substantial negative impact on downstream development activities: construction (i.e., hardware implementation – 77%) and testing (69%), and a moderate impact on the resultant system, with the physical design (24%) and reliability (18%) being the system quality attributes the most affected. Also, it was determined that the source of 70% of the constrained decisions was feedback from implementation activities.

3.3 Experiences

In this subsection, we describe the preliminary specific transitioning experiences we encountered when conducting the two studies.

3.3.1 Case study context

In examining the rows pertaining to this cluster (see Table 1), it is clear that the large-scale project context differs significantly from the lab study. The project scale, budget, domain complexity, development cycle time and team size of the RailCab project greatly exceed those of the lab study’s project. This type of context mirrors significant elements from the real-world industrial settings and alleviates the study threat from the lab study where students were used in an academic setting. However, because of the size of the project, and specifically the development time lifecycle, new challenges were encountered.

In the lab study, the project was contained within a two-month span and the study was conducted in accordance with the project itself. It was designed as a single iteration of a requirements project with an existing SA. The RailCab project had already been in development for 10 years, and the notion of iterations, due to the prototypical nature of the project, was not clear. Prior to the study, we had to determine a number of issues, such as: what parts of the development cycle should be measured? What data exists for this 10-year period? What constitutes an iteration in a prototypical development process? Can we “cleanly” separate and measure the RE and SA processes, as done in the lab study? Are requirements documented? Is the architecture documented? Are requirements incrementally added over time and the corresponding growth in system design be stratified as iterations?

The domain complexity between the projects in the studies was substantially different. The lab study’s banking domain was deliberately chosen as a familiar domain for the researchers and participants. This familiarity meant that the requirements work done in the projects, and the subsequent study analysis, was minimally affected by domain issues. However, the complexity of the innovative rail domain (e.g., advanced mechanical, electrical, control engineering, and integrated software issues), and heavy emphasis on “systems engineering” work rather than the lab’s “software engineering” focus, meant that the researchers would need to invest time and effort, and also require assistance from RailCab’s domain experts, in learning about the domain which, clearly, was critical for conducting the study. The challenge then became: how to familiarize with the complex domain in an unobtrusive way, in terms of RailCab staff’s involvement. To deal with this challenge, we prompted various staff members for internal and external documentation that could be researched off-line, and integrated “clarification interviews” within our study data collection interviews.

One advantage to working in the lab study is the higher number of projects (e.g., 6 cases in the [5] study) that is limited only by the number of students in the class and size of teams (e.g., 2/team in [5]) that are appropriate for the project. Having more cases leads to comparisons of results across projects that, although still not highly generalisable outside of the class project, at least more strongly indicate early trends and patterns that are emerging in the data. However, in a large-scale study, it is more likely that there is just one case to investigate, which can lead to generalizability issues (see 3.3.2). To mitigate this threat, we instead chose to execute the study separately on the major RailCab subsystems (e.g., energy management, drive and brake, active guidance, etc.) and report results at both the aggregate...
system level and at the individual subsystem level. This allowed us to identify and interpret results that were unique to one subsystem, and identify results that were consistent across the various subsystems. This analysis was possible because of the high complexity of each subsystem, and the low coupling among them; they each essentially could be considered as a separate system.

Another challenge when transitioning to the large-scale context is due to both the length of development time and the project staff turnover. The lab study, as seen in Table 1, had project duration of 2 months and also no participant turnover existed, meaning all participants could be contacted during or even after the study (until they left University). This facilitated meetings with all participants regarding every requirements decision made. However, in the RailCab project, many of the staff involved with earlier RE-SA decisions had long since quit the project and were not available for consultation, which led to us deciding that only the current up-to-date set of requirements decisions would be examined, and that we would only qualitatively investigate earlier decisions in terms of their effect on the current decisions where possible.

Though not a central issue in all replication projects, an interesting twist in our case was that the lab projects were all conducted in English; whereas, the RailCab project was conducted exclusively in German. With the core researchers’ working language being non-German made comprehension of the RailCab’s project documentation an added complexity of significance. The posed threat was contained by engaging another, bilingual, researcher who interpreted the German documents in English. Triangulating with the RailCab’s project staff (also bilingual) during interviews confirmed the accuracy of the interpretation and our comprehension of the technical issues. This added step in the RailCab study, certainly consumed significant human and time resources. We note that in international projects there can be multi-lingual issues to contend with.

3.3.2 Study design

With reference to Table 1, when examining the rows pertaining to the theme of Study Design, almost every facet of the study design for the RailCab study had to be altered in some way due to the changes in the context (described in the previous subsection).

The primary research question remained the same (see Q1, section 3.2.1), but we extended the lab study by adding two new research questions (see Q2 and Q3, section 3.2.2) that were not possible in the lab study; they required the full implementation of the project and the real-world context. These two additional questions led to the new constructs being investigated (impact on downstream activities and source of decision), though the main construct (requirements decisions) remained the same in both studies. This fulfills the basic requirement of replication, where at the very least the core research questions and constructs remain the same.

For the participants’ involvement, availability sampling [2] was used in the lab study; the researchers created the teams “artificially” to populate the projects. This was possible because of the high degree of researcher control in the study. In the RailCab study, however, the participant selection was much more influenced, none. In order to determine the quantitative results that were reported, the data was exported to an analysis tool and simple frequency counts were done on the various categories. The RailCab study analysis was similar, but required an extra major step before the above analysis could be done, and that is qualitative coding of the interview transcripts to create the categorization of requirements decisions. Furthermore, the same technique was used to qualitatively explore the two new research questions in the study.

The lab study, since it was conducted in an academic setting, had the obvious threats to validity when using students as participants. One of the main motivational points for conducting a large-scale case study is that these threats no longer apply. Despite avoiding these threats, the large-scale study brought on its own threats, many of which were not initially expected. In particular, threats to construct validity arose, since the measurement of the constructs was not as “clean” and required more researcher interpretation because data collection mechanisms (such as the data-logging tool in the lab study) could not be pursued.

4. Lessons Learnt and Recommendations

We describe below some example key lessons learnt.

1. Integrate domain knowledge acquisition as part of the case study process.

Not only were we scaling up in transitioning from the lab study to the prototypical study but we were also switching the application domain completely from banking to the rail system. With domain

---

1 Because this is a short paper we only have space to provide several examples.
switch, it is quite possible that the study investigators are not entirely familiar with the new domain. This was so in our case. Thus, we had a learning curve at the start of the study to understand the key aspects of the systems engineering project. Without this domain knowledge acquisition, we have no doubt that the case study was doomed to failure.

2: Lack of researcher control over the setup of the project under study means the researcher must adapt all investigative procedures to align with the case study context.

In the lab study, the researchers could carefully design and implement the study procedures to ensure quality of the study. However, in the large-scale study, the researchers did not have the means of “manipulating” the project details, and had to collect data without altering the project in any way. This led to different means of collecting and analyzing data, selecting participants, and introduced new threats to validity.

3: Case study projects can differ significantly from original lab study projects; thus it may be possible to extend the original study with new research questions by considering the case study environment.

At the outset of our large-scale study, we had intended to investigate only the research question (Q1) used in the lab study. However, it became clear quite early on upon preliminary discussions with key stakeholders that the breadth, depth, and richness of the RailCab case study data opened up new possibilities for investigative questions, in our case Q2 and Q3 (see Section 3.2.2.).

4: Researchers cannot expect to impose data collection mechanisms that (significantly) add time, effort, cost or burden to the development project in a large-scale study; data collection should instead focus on non-obtrusive and flexible schemes such as interviews, surveys and analysis of existing or emerging artefacts.2

Generally, project staff have no motivation or obligation to change their routines because of the case study project.

5: Avoiding lab study threats from the large-scale study does not mean that there would be no new or unexpected threats in the latter study.

Often, case study threats can arise because the measurement of the constructs is not as “clean” as in the lab study and can require more researcher interpretation due to lack of automation of data gathering.

5. Conclusions and Future Work
Repetition is considered important in scientific research in SE [4]. However, the number of published repetition studies is quite low [6], suggesting a rather meager body of knowledge pertaining to this topic. In particular, significant knowledge on “transitioning” from lab studies to large-scale case studies is nonexistent. We conducted a large-scale replicated case study [3] that extended an original laboratory study [5] investigating the

“effect an existing systems architecture has on requirements decisions”.

In this short paper, we describe the preliminary analysis of our researcher experiences when transitioning from the lab study to the large-scale case study. Our analysis centered on numerous dimensions, including differences in the case study context (e.g., project scale, budget, domain, etc.) and study design (e.g., data collection procedures, threats to validity, research questions, etc.).

Our main conclusion is that it is generally not possible to retain the original lab study design for a large-scale case study. Due to the increase in scale of the project and its context, and the lack of researcher control over the project and its environment, nearly every facet of the study design must be adapted. Beyond the basic research question(s) and constructs, our replicated study featured an almost new design. Further to this general conclusion, the paper also provides lessons learnt regarding specific aspects of the transitioning between the two studies. Such lessons learnt represent “do’s and don’ts” that are important for other researchers wishing to extend their preliminary lab studies into industrial settings. This can lead to more transitional studies being conducted by other researchers, which ultimately can lead to SE practice more likely to use empirical results in driving software development and maintenance [4]. Our ongoing work involves analyzing our experiences from further study attributes.

6. Acknowledgements
This work was, in part, supported by Natural Science and Engineering Research Council (NSERC) of Canada. This contribution was also developed and published in the course of the Collaborative Research Center 614 “Self-Optimizing Concepts and Structures in Mechanical Engineering” funded by the German Research Foundation (DFG) under grant number SFB 614.

7. REFERENCES


2 Another avenue is automated data collection, but this was not a possibility in our study.