Joint Analysis of Families of SE Experiments

Adrián Santos Parrilla
University of Oulu, Finland
Department of Information Processing
Science
adrian.santos.parrilla@oulu.fi

ABSTRACT

Context: Replication is of paramount importance for building solid theories in experimental disciplines and is a cornerstone of the evolution of science. Over the last few years, the role of replication in software engineering (SE), families of experiments and the need to aggregate the results of groups of experiments have attracted special attention. Frameworks, taxonomies, processes, recommendations and guidelines for reporting replications have been proposed to support the replication of SE experiments. There has been much less debate about the issue of the joint analysis of replications whose raw data are available to experimenters.

Objectives: The aim of our research is to explore current trends in the joint analysis of SE experiments whose raw data are available to experimenters. Notice that the fact that experimenters have access to the raw data is what differentiates joint analysis from other methods for aggregating experimental results (e.g. systematic literature review (SLR), where the applicability of meta-analysis techniques is widely accepted). The objective of this three-year investigation is to shed light on the best joint analysis approach when the experimenters have access to raw data from several replications.

Method: Narrative comparison, standard frequentist methods, meta-analysis and Bayesian methods have been used in SE literature. We will apply and evaluate each approach to the experiments on Test-Driven Development (TDD) carried out within the Experimental Software Engineering Industrial Laboratory (ESEIL) project. We will propose and rate a tentative framework for aggregating results within the ESEIL project. The proposed framework, as well as the different existing methods, will be evaluated on another set of replications of testing technique experiments.

Current status: The thesis proposal was elicited on the 15 January 2015 and rounded out over the following six months. As a three-year thesis, its discussion and findings will be projected across the years 2015, 2016 and 2017. The first results are now being aggregated with the data from four different experiments on TDD (two in academia and two in industry), and preliminary results are expected to be available in October 2015.

Keywords

SE replication, joint analysis, family of experiments, raw data.

1. INTRODUCTION

SE experiments can be analyzed separately to acquire knowledge about the performance of different treatments under certain circumstances (working environment or specific population characteristics). The shortcomings of this approach include: (1) the number of subjects is a limiting factor across most SE experiments; (2) the results might be artifactual, that is, due to the impact of the experimental protocol and not to the treatments applied by the subjects; (3) the findings from one study cannot be interpreted outside the confines of the setting of that experiment.

The role and importance of replications in tackling the issue of the generalization of SE experimental findings has been recognized by many authors within the SE community [11, 19, 22, 24]. As stated in [24], “replications play a key role in Empirical Software Engineering by allowing the community to build knowledge about which results or observations hold under which conditions”. The aim of replication is twofold [23]. First, “replication is needed not merely to validate one’s findings, but, more importantly, to establish the increasing range of radically different conditions under which the findings hold, and the predictable exceptions”. Second, as noted in [11], “if an experiment is not replicated, there is no way to distinguish whether results were produced by chance (the observed event occurred accidentally), results are artifactual (the event occurred because of the experimental configuration but does not exist in reality) or results conform to patterns existing in reality”. Thus, replication provides experimenters and the community with a continuous knowledge building process by: (1) confirming previous experimental results; and (2) identifying the reasons why previous results do not hold under the new experimental conditions. By aggregating the results of experiments, we get to see the whole picture for different population characteristics, settings and conformance to the treatments used within the SE community.

The shortage of replication studies within the SE community was highlighted in [28]. Out of a total of 5453 articles published in different SE-related journals and conference proceedings between 1993 and 2002, 20 out of 113 controlled experiments were described as replications [28]. In a mapping study on SE replications completed from 2010 to 2011 [9] based on bibliographic searches covering the period from 1994 to 2010, only 96 out of 16,000 papers included replications. These 96 papers reported a total of 133 replications. Furthermore, the results showed that nearly 70% of the replications were published after 2004 and that up to a 70% of the studies were *internal* replications (i.e., carried out by the same experimenters) [9]. An update of the same study identified and analyzed replications published in 2011 and 2012, and noted that the trend in the number of replications in SE continued to be upward (56 papers in two years) [9]. However, the growth rate was slow, possibly indicating the need for patterns to improve the way in which replications are run in the field [9].

The organization of an International Workshop on Replication in Empirical Software Engineering Research (RESER) is illustrative of the growing interest in replication. At this venue, empirical software engineering researchers have the opportunity to present and discuss the theoretical foundations and methods of replication, as well as the results of replicated studies.

Traditionally groups of experiments have been formed within the SE community by means of SLR, and their results analyzed jointly in order to build new pieces of knowledge. But, nowadays, researchers are replicating their own studies in order to increase the relevance and validity of their findings.

As some authors state [29], there is a need to further investigate the problem of generalizing conclusions from individual studies. This could be done by extending research tools commonly used in engineering and computer science with those applied in sciences that study people such as medicine or psychology. Brooks [5] suggested that research methods like statistical meta-analysis could benefit software engineering in generalizing the findings from individual studies. However SE experiments have in general several constraints which make difficult the application of meta-analysis [10]: (1) small sample sizes (generally less than 10 subjects per treatment); (2) the number of experiments per meta-analysis is also small in many cases; (3) some studies do not provide the statistical parameters required for meta-analysis when reporting their results.

Even though meta-analysis is a widely accepted method for aggregating results from studies identified by means of SLR (generally reporting statistics such as the mean, standard deviation or number of subjects), there appears to be no such agreement on the right way to analyze the replications of experiments whose raw data are available to experimenters. Researchers who are in possession of the raw data of the experiments are better able to compute, understand and assess the different variables considered in the experiments than if they only have access to findings reported in different publications. Furthermore, the issue nowadays seems to be object of debate in other fields such as medicine or social sciences where the communities are still discussing the advantages and disadvantages of conducting meta-analysis with individual participant data (IPD) gathered from the constituent studies and aggregated data (AD), or the group-level statistics (effect sizes) that appear in reports of a study's results [7].

It is unclear yet within the SE community which is the most straightforward and valid procedure for aggregating results when the raw data of the experiments are available to the experimenters and they have first-hand knowledge of the protocol and conditions. Besides, different joint analysis techniques may be applicable depending on the different characteristics of the replications.

The concept of family of experiments was first reported in SE by Basili et al. in 1991 [3]. This concept is explained in [1] as follows: "a family is composed of multiple similar experiments that pursue the same goal to build the knowledge needed to extract significant conclusions". From this point of view, the concept of family of experiments is a "framework for organizing sets of related studies" [3], where "experiments can be viewed as part of common families of studies rather than being isolated events" [3]. As each of these experiments is viewed as belonging to a group of studies, their results could be analyzed as a whole instead of separately, and the findings integrated into one comprehensive result.

Notice that this definition of family of experiments also covers related experiments found by a SLR, even though the experimenters are completely unconnected. We think that the concept of family of experiments should be defined more precisely in order to make a distinction between the two situations below:

i. Set of related experiments, typically found in a SLR, that can be aggregated to generate evidence. The available information in these cases provides only a short description of protocol and conditions as regards the setting and no more than sample descriptive statistics as regards the data.

ii. Set of experiments conducted by related researchers that make the raw data available for further joint analysis. In this case, the available information covers everything that the experimenters know about their own studies.

We suggest that the term family of experiments should be used to refer to situation (ii) above. Thus, this research narrows down the meaning of family of experiments to a definition similar to the explanation given in [1]: "a set of similar experiments that pursue the same goal to build the knowledge needed to extract significant conclusions", where experimenters have the raw data of the experiments and first-hand knowledge of the setting.

In this article we report a PhD thesis that is being carried out to investigate how to conduct a joint analysis of a family of experiments. We first report the current methods that have been used in SE for the joint analysis of experiments. We also outline a tentative path for building a framework for aggregating the results of families of experiments.

This paper is organized as follows. Section 2 briefly discusses relevant prior work on the topic of results aggregation in SE. Section 3 outlines the main objectives of the proposal. Section 4 describes the proposed research approach. Finally, Section 5 summarizes the current status of the outlined proposal.

2. RELATED WORK

In the following sections, we briefly discuss the different approaches proposed and adopted within the SE community to conduct joint analyses of families of experiments whose raw data are available to experimenters, discuss their applicability and state the conclusions concerning their use reported in the different publications.

The different techniques are discussed in chronological order by date of publication of the respective paper applying or proposing the technique.

2.1 Narrative Comparison

The difference between close and differentiated replications was discussed in depth by Juristo and Vegas in 2011 [19]. They proposed an approach for analyzing groups of experiments: the results of each experiment are analyzed separately and are then grouped according to concordances and discordances between the results of the replications identified through narrative comparison. A differentiated replication (i.e., a replication that produces a different outcome than the main experiment) is considered as an opportunity to explore the different variables that might have had an impact on the outcome rather than being seen as a threat to the validity of the replication. There are several noteworthy points with regard to the study reported in [19]:

- There is a big imbalance between the total number of subjects participating in each replication (176 participants at the UPM; 31 subjects in the UPV replication and 76 in the ORT replication). This imbalance in the number of subjects could have biased the results due to natural random variability.
- The report states that "the results are considered equal if the estimated mean value for the replication results is within the confidence interval of the baseline experiment results" [19]. Some sources [8] state that roughly 83% of replications will fall within the 95% interval of confidence for the means of the original experiment. In other words, even if two samples (one per experiment) are drawn from the same population, there is an 83% chance that the mean of the second experiment will fall within the confidence interval of the first. Thus, due to the random variability of the sample, Juristo and Vegas might be considering the result of the replication as a different outcome, merely because the mean did not fall within the confidence interval of the main experiment (although, in actual fact, it represents the same result, i.e., population, in a different random sample). This underestimation of sampling variability is a limitation of the applicability of the narrative comparison approach proposed in [19]: roughly 17% of replication results will, due to random variability, be considered different when they are in fact equal (i.e., represent the same population).

This method relies on narrative comparison to analyze a family of experiments using observations such as the mean of the different outcome variables obtained by the subjects in the experiments to discuss the findings. It is up to the expert analyst to observe and interpret the outcomes, and the method depends on their ability to identify extraneous factors that might have influenced the outcome in the different replications. One clear drawback is that the technique might underestimate the random sampling variability within a certain population and thus overestimate the effect of third variables. This technique can be applied to the raw data of the experiments or to the known descriptive statistics of the different experimental outcomes (although the variables should be interpreted with due caution if the raw data are not available).

In any case, the narrative comparison technique requires the experimenters of the replications to interact in order to identify extraneous variables that might have had an impact on the consistency of results in order to gather knowledge for further investigation.

2.2 Meta-Analysis

Meta-analysis is a set of statistical techniques that has been used to combine the different effect sizes of a family of experiments [9, 10]. Effect sizes can be estimated to evaluate the average impact of an independent variable on a dependent variable across studies. Since measures may be taken from different settings and may be non-uniform, a standardized measure must be taken for each experiment. These measures must be combined to estimate the global effect size of a factor [1].

Meta-analysis is the current standard for aggregating quantitative results across studies [22]. It can be used to combine data even if studies report contradictory results provided that the overall variation is not too extreme [25].

Meta-analysis has been used within the SE community with multiple objectives such as studying the effects of TDD on external quality and productivity [26], checking for correlations of metrics across software projects corpora [30] or the effect on defect detection rates of different inspection techniques [14] amongst others [6, 13]. Furthermore, guidelines for applying different meta-analytic techniques have been proposed to the SE community [10].

Traditional meta-analytic techniques rely on the assumption that effect size estimates from different experiments are independent and have sampling distributions with known conditional variances [16]. An experiment that examines multiple dependent variables or a cluster of studies carried out by the same investigator or laboratory [16] poses a threat to the supposed independence of experiments. The hierarchical dependence model is applicable when the dependence structure between the experiments is due to the inherent condition of belonging to a cluster of experiments [12]. The circumstances, implications and impact of dependence across studies have been studied at length by different researchers [15, 21], and multiple techniques for dealing with this issue have been proposed [16]. However, their usage requires an in-depth knowledge of the different techniques available, and their applicability is by no means clear [16].

How meta-analysis should be applied to a family of experiments in SE is a matter of debate, and there are many opinions on procedure. As stated by Miller [24], "because the dependent replications rely on the same underlying protocols as the original study, their results cannot be considered as truly independent of the original study. Moreover, they may propagate any accidental biases from the original study into the results of the replication". Recall, which is one of the main assumptions underlying traditional meta-analysis, relies on independence and could be violated in some cases where replications are run by related researchers.

Again, if experimental material is reused (thus increasing the dependence between two experiments), "although from a simple replication point of view, this seems attractive; from a meta-analysis point of view this is undesirable, as it creates strong correlations between the two studies" [24]. Kitchenham shares this view [22], stating "in particular, dependent replications violate the main assumption underlying meta-analysis which is the standard method of aggregating results from quantitative experiments. Recently, my colleagues and I were forced to omit three studies from a systematic literature review because the 'replications' were so close that they offered no additional information to the aggregation process".

Pickard et al. [25] state, in reference to the outcome of the primary studies, that "the greater the degree of similarity between the studies the more confidence you can have in the results of a meta-analysis". The best thing then would be very similar settings without either communication or information sharing among experimenters: a rare occurrence in SE.

Furthermore, it is up to researchers to settle several issues regarding meta-analysis, such as:

- The selection of the effect size metric used to perform the joint analysis, i.e., computation of the raw mean difference, a standardized mean difference, odds ratio, risk ratio or risk differences [4].

- The standardizer used to compute the effect size, i.e., pool standard deviations, weight each group's standard deviation by sample size or use the control group standard deviation [8].
- The computation of some effect sizes from others or the use of unbiased versions of effect size metrics [8].

Besides, the impact of experimental designs on the resulting effect sizes (such as multiple-treatment studies and multiple-endpoint studies [16]) makes meta-analysis applicability a controversial topic in the SE community.

Meta-analysis has the potential of aggregating the results of different experiments if the raw data are not available, even though its stability in SE experiments has been questioned [24]. However, meta-analysis has been applied as well when the experimenters are in possession of the raw data [1]. This issue raises the question of whether the best procedure for the joint analysis of experiments whose raw data are available to researchers is to apply meta-analysis techniques or whether it would be better to use other approaches. As noted in [4]: "losing sight of the fact that meta-analysis is a tool with multiple applications causes confusion and leads to pointless discussions about what is the right way to perform a research synthesis, when there is no single right way".

All the above raises doubts within the community, which does not appear to be clear about the applicability of meta-analysis, its boundaries and misuses, adding to the confusion surrounding the aggregation of results in families of SE experiments.

2.3 Standard Frequentist Methods

Another option for conducting the joint analysis of families of experiments is to analyze all data together via standard frequentist methods such as analysis of variance (ANOVA) [27]. Basically, each experiment within a group of replicated controlled experiments is analyzed separately. After gathering knowledge about the results of the different replications and briefly discussing whether or not the results hold, the experimenters hypothesize about which variables might have had an impact on the results. In a next step, the raw data from all the different studies belonging to the family of experiments are aggregated and analyzed as a whole considering the hypothesized variables as factors.

The study presented by Runeson et al. [27] in 2014 is an example of such an approach. They report three experiments comparing code inspections with unit testing: the original experiment, an internal replication (a replication performed by the same researchers minimizing changes in the replication) and an external replication (a replication performed by a different group of researchers, varying several aspects of the experiment) [27]. The three experiments were cross-over designs, where the subjects applied one defect detection method (code inspection or structural unit testing) to one program and then the other method to the other program [27]. The dependent variables of the experiments are time spent on the tasks, number of defects detected and localized and rate, i.e., number of defects detected and localized per time unit [27].

The separate analyses performed for each experiment in [27] appear to be clearly explained from the data analysis viewpoint: "the experiment has two factors, paired measurements, a sample size of less than 30 and data which is not normally distributed".

When aggregating the data from the three experiments into one data set and carrying out the joint analysis, however, Runeson et al. report [27] "the overall two-factor ANOVA results for the three experiments". Notice that the authors no longer mention that the data are "repeated measures", and the "two-factor ANOVA" analysis carried out is interpreted without any reference to a within-subjects factor. Furthermore, the joint analysis of the three experiments is performed using a Kruskall-Wallis test, the equivalent of the one-way ANOVA test for non-normal distributions.

Because the data are dependent, a repeated measures general linear model could have been fitted to analyze the data. Also, the within-subjects and between-subjects factors considered should have been clearly specified in order to pave the way for data analysis, understandability and reproduction.

A framework applying this approach to aggregate results from a family of experiments should comply with three objectives: (1) provide a specific set of steps to be carried out to pre-process the data and report the data pre-processing of the experiments; (2) provide a template with all the relevant information that should be stated about each of the experiments to carry out the individual analysis; (3) provide guidance for defining a joint analysis from separate experiments, accounting for any of the possible limitations of each experiment.

2.4 Bayesian Methods

Bayesian methods for data analysis have also been applied to the aggregation of the results of a family of experiments [17, 18]. They resolve the inconsistencies found between replications and the original experiment by investigating moderators, i.e., variables that cause an effect to differ across contexts [2]. An iterative approach is applied to try to identify moderators that might have an influence on the outcome of the experiment. The different variables and their interaction studied in the proposed models are then measured based on the most relevant changes made to the different replications. As explained in [17], "By moderator, we mean any explanatory variable that interacts with another explanatory variable in predicting a response variable. For one variable to "moderate" another does not mean that it dampens the other's effect —rather, it means that an interaction exists, such that the latter's effect varies in response to the former".

Bayesian methods provide an alternative to traditional meta-analysis. First, using Bayesian methods, data can be accumulated over time (prior knowledge) into the analysis of future replications. Second, Bayesian methods can be used to combine results such that all data are treated as current observations [18].

However, the application of this method to joint analysis requires thorough knowledge of Bayesian statistics, which have seldom been used in the SE community that is dominated by frequentist methods such as meta-analysis: the current standard for aggregating quantitative results [22, 25].

3. RESEARCH OBJECTIVES

We have carried out several experiments to assess the TDD agile development technique [31]. A lot of data are being collected from multiple replications, and their analysis and processing could provide insights into proper ways of handling and synthesizing the results of joint analysis.

Our research is driven by several methodological questions:

1. What is the best way of analyzing families of experiments with raw data in SE? Do the existing approaches produce contradictory results? Under what circumstances are these different analysis approaches applicable?
2. Where are the limits to the *feasibility* of grouping different experiments, i.e., how similar does the design of the experiment need to be?
3. Is there any kind of knowledge on the different experiments that is of paramount importance for joint analysis? Is it always correctly reported?

Also, several TDD-specific questions will drive our research:

1. Does subject type (students or professionals) have any implications regarding the performance of different development techniques (TDD, ITL)?
2. Does any moderator variable or interaction amongst moderator variables across different organizational and/or academic setups have an impact on the performance of the development methodologies?
3. Does the context (academia, industry or even different industries) have any impact on the performance of the subjects applying different development techniques?

Other research questions might arise in the course of the research, and their implications will be discussed thoroughly as part of the PhD thesis.

Our research will provide different contributions to the academia and practice:

1. Different methods for aggregation of results will be used jointly for analyzing families of experiments whose raw data is available to experimenters, and the edges of applicability of the different techniques will be discussed along the research process.
2. Multiple industry experiments on TDD will be aggregated and their results for different software metrics analyzed. Specifically, different treatments such as traditional test last coding or ITL will be compared against TDD in industrial settings, which may lead to interesting findings about the effects of TDD in real software development contexts.

4. RESEARCH APPROACH

Within the ESEIL project, several replications (i.e., reporting both different and consistent results) are being run on the topic of TDD performance. These replications consistently alter different aspects of the primary study (design of the experiment, subject type, instrumentation, treatments, artifacts, location, training, researchers, session length, etc.). All these replications are being analyzed separately and their findings discussed.

In order to carry out the joint analysis of experiments, we will first run a search of current trends in the analysis of families of experiments in other areas such as agriculture, psychology or medicine. Such bibliographic searches of online databases could turn up a variety of methods or prescriptions that might lend themselves to extrapolation to SE. The conditions under which these techniques can be used and their limitations will be studied within specific SE setups, and they will be assessed by means of direct application to the ESEIL project experiments on TDD.

The application of these different analysis techniques to the same family of experiments can lead to multiple, possibly even contradictory results. In our studies we will try to explore the scope of application of the different aggregation approaches and define the limits of their applicability for conducting joint analysis.

After exploring these approaches and discussing their implications within the ESEIL project, we will propose a framework for the joint analysis of families of experiments. A second version of this framework will be refined and further expanded with the aim of applying it to a family of testing experiments [20].

We will then adopt the most promising methods in order to extrapolate their applicability to a different set of experiments within the SE community such as software requirements.

Finally, the proposed updated framework could be assessed and reviewed by colleagues within the SE community from different viewpoints in order to lend the proposal higher external validity and consistency.

5. SUMMARY OF CURRENT STATUS

The thesis proposal was elicited on 15 January 2015 and rounded out over the following six months. As a three-year thesis, its findings and proposals will be projected across the years 2015, 2016 and 2017. The publishing strategy targets publication at the ICSE and ESE conferences and in the TSE, TOSEM, EMSE and IST journals over the three-year research period.

At the time of writing, a preliminary aggregation of results is being carried out using the data from four replications as part of the Experimental Software Engineering Industrial Laboratory (ESEIL) project. Two experiments were run in a professional setting, whereas another two were run in academia. The results of the experiments will be aggregated using different analysis approaches, and their implications, constraints and findings will be discussed and further explored in subsequent studies.

6. REFERENCES

[1] Abrahao, S., Gravino, C., Insfran, E., Scanniello, G., & Tortora, G. (2013). Assessing the effectiveness of sequence diagrams in the comprehension of functional requirements: Results from a family of five experiments. *Software Engineering, IEEE Transactions on*, *39*(3), 327-342.

[2] Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, *51*(6), 1173.

[3] Basili, V. R., Shull, F., & Lanubile, F. (1999). Building knowledge through families of experiments. *Software Engineering, IEEE Transactions on*, *25*(4), 456-473.

[4] Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011).*Introduction to Meta-Analysis*. John Wiley & Sons.

[5] Brooks, A. (1997). Meta analysis—a silver bullet—for meta-analysts. *Empirical Software Engineering*, *2*(4), 333-338.

[6] Ciolkowski, M. (2009, October). What do we know about perspective-based reading? An approach for quantitative aggregation in software engineering. In*Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement* (pp. 133-144). IEEE Computer Society.

[7] Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data.*Psychological methods*, *14*(2), 165.

[8] Cumming, G. (2012). Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. Routledge.

[9] Da Silva, F. Q., Suassuna, M., França, A. C. C., Grubb, A. M., Gouveia, T. B., Monteiro, C. V., & dos Santos, I. E. (2014). Replication of empirical studies in software engineering research: a systematic mapping study. *Empirical Software Engineering*, *19*(3), 501-557.

[10] Dieste, O., Fernández, E., Garcia Martinez, R., & Juristo, N. (2011, April). Comparative analysis of meta-analysis methods: when to use which?. In*Evaluation & Assessment in Software Engineering (EASE 2011), 15th Annual Conference on* (pp. 36-45). IET.

[11] Gómez, O. S., Juristo, N., & Vegas, S. (2014). Understanding replication of experiments in software engineering: A classification. *Information and Software Technology*, *56*(8), 1033-1048.

[12] Gurevitch, J., & Hedges, L. V. (1999). Statistical issues in ecological meta-analyses. *Ecology*, *80*(4), 1142-1149.

[13] Hannay, J. E., Dybå, T., Arisholm, E., & Sjøberg, D. I. (2009). The effectiveness of pair programming: A meta-analysis. *Information and Software Technology*, *51*(7), 1110-1122.

[14] Hayes, W. (1999). Research synthesis in software engineering: a case for meta-analysis. In *Software Metrics Symposium, 1999. Proceedings. Sixth International* (pp. 143-151). IEEE.

[15] Hedges, L. V., & Olkin, I. (2014). *Statistical method for meta-analysis*. Academic press.

[16] Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39-65.

[17] Jonathan L. Krein, Lutz Prechelt, Natalia Juristo, Aziz Nanthaamornphong, Jeffrey C. Carver, Sira Vegas, Charles D. Knutson, Kevin D. Seppi and Dennis L. Egget, A Multi-site Joint Replication of a Design Patterns Experiment using Moderator Variables to Generalize across Contexts. *Software Engineering, IEEE Transactions.* Under review.

[18] Jonathan L. Krein, Lutz Prechelt, Natalia Juristo, Kevin D. Seppi, Aziz Nanthaamornphong, Jeffrey C. Carver, Sira Vegas and Charles D. Knutson, A Method for Generalizing across Contexts in Software Engineering Experiments. *Software Engineering, IEEE Transactions.* Submitted.

[19] Juristo, N., & Vegas, S. (2011). The role of non-exact replications in software engineering experiments. *Empirical Software Engineering*, *16*(3), 295-324.

[20] Juristo, N., Vegas, S., Solari, M., Abrahao, S., & Ramos, I. (2012, April). Comparing the effectiveness of equivalence partitioning, branch testing and code reading by stepwise abstraction applied by subjects. In *Software Testing, Verification and Validation (ICST), 2012 IEEE Fifth International Conference on*(pp. 330-339). IEEE.

[21] Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological methods*, *1*(3), 227.

[22] Kitchenham, B. (2008). The role of replications in empirical software engineering—a word of warning. *Empirical Software Engineering*, *13*(2), 219-221.

[23] Lindsay, R. M., & Ehrenberg, A. S. (1993). The design of replicated studies. *The American Statistician*, *47*(3), 217-228.

[24] Miller, J. (2000). Applying meta-analytical procedures to software engineering experiments. *Journal of Systems and Software*, *54*(1), 29-39.

[25] Pickard, L. M., Kitchenham, B. A., & Jones, P. W. (1998). Combining empirical results in software engineering. *Information and software technology*, *40*(14), 811-821.

[26] Rafique, Y., & Misic, V. (2013). The effects of test-driven development on external quality and productivity: A meta-analysis. *Software Engineering, IEEE Transactions on*, *39*(6), 835-856.

[27] Runeson, P., Stefik, A., & Andrews, A. (2014). Variation factors in the design and analysis of replicated controlled experiments. *Empirical Software Engineering*, *19*(6), 1781-1808.

[28] Sjøberg, D. I., Hannay, J. E., Hansen, O., Kampenes, V. B., Karahasanovic, A., Liborg, N. K., & Rekdal, A. C. (2005). A survey of controlled experiments in software engineering. *Software Engineering, IEEE Transactions on*, *31*(9), 733-753.

[29] Succi, G., Spasojevic, R., Hayes, J. J., Smith, M. R., & Pedrycz, W. (2000). Application of statistical meta-analysis to software engineering metrics data. In*Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics* (Vol. 1, pp. 709-714).

[30] Succi, G., Spasojevic, R., Hayes, J. J., Smith, M. R., & Pedrycz, W. (2000). Application of statistical meta-analysis to software engineering metrics data. In*Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics* (Vol. 1, pp. 709-714).

[31] Vegas, Sira; Dieste, Oscar; Juristo, Natalia, "Difficulties in Running Experiments in the Software Industry: Experiences from the Trenches," Conducting Empirical Studies in Industry (CESI), 2015 IEEE/ACM 3rd International Workshop on , vol., no., pp.3,9, 18-18 May 2015 doi: 10.1109/CESI.2015.8