Validity Threats in Empirical Software Engineering Research - An Initial Survey

Robert Feldt, Ana Magazinius
Dept. of Computer Science and Engineering
Chalmers University of Technology
SE-412 96 Gothenburg, Sweden
Email: robert.feldt@chalmers.se

***Abstract*—In judging the quality of a research study it is very important to consider threats to the validity of the study and the results. This is particularly important for empirical research where there is often a multitude of possible threats. With a growing focus on empirical research methods in software engineering it is important that there is a consensus in the community on this importance, that validity analysis is done by every researcher and that there is common terminology and support on how to do and report it. Even though there are previous relevant results they have primarily focused on quantitative research methods and in particular experiments. Here we look at the existing advice and guidelines and then perform a review of 43 papers published in the ESEM conference in 2009 and analyse the validity analysis they include and which threats and strategies for overcoming them that were given by the authors. Based on this analysis we then discuss what is working well and less well in validity analysis of empirical software engineering research and present recommendations on how to better support validity analysis in the future.**

***Keywords*-Validity threats, Empirical research, Software engineering, Research methodology**

I. Introduction

Empirical research in software engineering (ESE) is increasingly important to advance the state-of-the-art in a scientific manner [1]. A critical element of any empirical research study is to analyze and mitigate threats to the validity of the results. A number of summaries, models and lists of validity threats (VTs) have been presented to help a researcher in analyzing validity and mitigate threats [1, 2]. Many of these results focus on VT analysis for quantitative research such as experiments [1] while other consider VT analysis for qualitative research [2].

For a researcher in ESE it may not be clear if the existing checklists are consistent or which one applies to a particular study. In particular for qualitative ESE research this can be a problem since few studies with results specific to software engineering has been presented. Furthermore, several recent results have pointed to problems in how ESE research is conducted or reported in sub-areas [3, 4, 5]. In this paper, we want to analyse in particular how validity threats are analysed and mitigated in ESE research studies. Ultimately this will lead to additional support and guidelines for how to more consistently perform such analysis and mitigation and thus increase the validity of future ESE results.

In this study we survey how validity threats are analyzed in recent papers published in ESE and discuss the current state-of-the-art and what may have caused it. We then propose ways that can help in improving the analysis of validity threats. In particular we address the research questions:

- RQ1. What are the existing advice on and checklists for VTs in ESE?
- RQ2. What is the current practice in discussing and handling validity threats in research studies in ESE?
- RQ3. Is the current practice of sufficient quality and, if not, how can it be improved?

To limit the scope the survey presented here has been constrained to only cover one year of a major ESE publication venue. Because of this our results are only preliminary.

Section II below describes existing advice on validity threats in software engineering. In Section III we describe the methodology for the survey we have performed on VT analysis and mitigation in ESE. Section IV gives the results of the survey and analyses them. Finally, Section V discusses the VTs of this study itself and future work, while Section VI concludes.

II. Validity in Software Engineering Research

Validity of research is concerned with the question of how the conclusions might be wrong, i.e. the relationship between conclusions and reality [2]. This is distinct from the larger question of how a piece of research might have low quality since quality has more aspects than validity alone, for example relevance and replicability. In non-technical terms validity is concerned with 'How the results might be wrong?' not with the larger questions of 'How this research might be bad?'. In many cases they overlap, though.

Validity is a goal, not something that can be proven or assured with the use of specific procedures. A *validity threat* (VT) is a specific way in which you might be wrong [2]. In a *validity analysis* (VA) you identify possible threats and discuss and decide how to address them. A specific choice or action used to increase validity by addressing a specific threat we call a *mitigation strategy*.

For quantitative research in software engineering, such as experiments, specific advice on validity analysis and threats was given by Wohlin et al [1] structured according to previous

results from Cook et al [6]. Wohlin et al discuss four main types of validity threats: conclusion, internal, construct and external.

Conclusion validity focus on how sure we can be that the treatment we used in an experiment really is related to the actual outcome we observed. Typically this concerns if there is a statistically significant effect on the outcome.

If there is a statistically significant relationship the *Internal validity* focus on how sure we can be that the treatment actually caused the outcome. There can be other factors that have caused the outcome, factors that we do not have control over or have not measured.

Construct validity focus on the relation between the theory behind the experiment and the observation(s). Even if we have established that there is a casual relationship between the treatment of our experiment and the observed outcome, the treatment might not correspond to the cause we think we have controlled and altered. Similarly, the observed outcome might not correspond to the effect we think we are measuring.

Finally, the *External validity* is concerned with whether we can generalize the results outside the scope of our study. Even if we have established a statistically significant casual relation between a treatment and an outcome and they correspond to the casue and effect we set out to investigate the results are of little use if the cause and effect we have established does not hold in other situations.

For qualitative research methods there are not software engineering specific results and we have to turn to other fields of study. Epistemologically qualitative research stretches between positivism as one extreme and interpretivism as the other [7]. The two similar notions of subtle realism and anti-realism have also been used by some authors [8].

The fundamental difference between the two resides in the interepretivists belief that the humans differ from the subjects/objects of study in the natural sciences since social reality has a meaning for humans. That meaning is the basis both for human actions and the meanings that humans attribute to their own actions as well as the actions of others [7]. Positivists on the other hand view humans as any other data source that can be sensed, measured and positively verified [7].

Both positivistic and interpretivistic scientists are interested in assessing whether they are observing, measuring or identifying what they think and say they are [9]. However, since their research questions, methods and views on reality differ, so do the methods to assess the quality of their work. Positivists usually use natural sciences criteria: internal and external validity, reliability and objectivity [7]. The interpretivists use a set of alternative criteria: credibility, transferability, dependability and confirmability [10]. No matter the criteria, there are threats to the validity of the research. Lincoln and Guba suggest reactivity (researchers presence might influence the setting and subjects behaviour), respondent bias (where subjects knowingly or unknowingly provide invalid information) and researcher bias (assumptions and personal beliefs of the researcher might affect the study) as main threats. Maxwell focuses on the researcher and lists description (of the collected data), interpretation (of the collected data) and theory (not considering alternative explanations) as the main threats [2].

Table I below summarizes the main validity threats that have been discussed above.

III. Review Methodology

Below we describe how we selected the papers to be included in the review, the overall process used and the forms used in the analysis of each paper.

A. Selection of papers

To limit the scope of our initial review we wanted to select only recent papers published in ESE. This is also motivated by the fact that ESE has been maturing and if we would include too many years back these papers might not be representative of current practice. Based on this we selected all full research papers published in the Empirical Software Engineering and Measurement (ESEM) conference in 2009. Even though this is a limited selection our assumption is that ESEM is the state-of-the-art conference on empirical software engineering research and as such can be expected to have higher expectations, standards and experience of empirical research.

A total of 43 full papers was published in ESEM 2009 and are thus included in our results and analysis below.

B. Process for Review

For each of the papers we read and filled in one form summarizing the type of and empirical content of the paper (Paper form). For each validity threat or issue discussed in the paper we then filled in a form with detailed information about each threat and any mitigation strategies mentioned for it (VT form). The forms used are presented in Tables II and III respectively and described in more detail below.

The forms themselves were created in discussions between the researchers based on the review of the existing literature on validity threats and analysis presented above. These initial forms were then piloted on a random selection of 4 papers which was reviewed independently by both authors. The pilot test did not uncover any discrepancies between the analyses of the two authors. It only prompted clarification of where to record what and the exclusion of a redundant question. Since we used the same papers in this step this also validated that we had the same interpretation of the forms and used them in the same way. We also discarded the idea of classifying each identified threat since there is no clear unified model for how this can be done.

The remaining 39 papers was then reviewed by one of the authors and the corresponding forms filled in. During this

Table I: Main types of Validity threats discussed in the literature.

Validity threat type	Example of typical questions to be answered
Conclusion validity	Does the treatment/change we introduced have a statistically significant effect on the outcome we measure?
Internal validity	Did the treatment/change we introduced cause the effect on the outcome? Can other factors also have had an effect?
Construct validity	Does the treatment correspond to the actual cause we are interested in? Does the outcome correspond to the effect we are interested in?
External validity, Transferability	Is the cause and effect relationship we have shown valid in other situations? Can we generalize our results? Do the results apply in other contexts?
Credibility	Are we confident that the findings are true? Why?
Dependability	Are the findings consistent? Can they be repeated?
Confirmalibity	Are the findings shaped by the respondents and not by the researcher?

review work, whenever some aspect of a paper were unclear the paper were discussed jointly by the authors.

In total this resulted in 43 Paper forms and a total of 136 VT Issue forms filled out. One paper is excluded from our analysis since it was a theoretical paper without any empirical content or any analysis of validity. Descriptive statistics, aggregation and analysis of the forms was then done cooperatively between the researchers.

C. Paper form

Table II presents the paper form that was filled out for each of the reviewed papers.

The questions on the form aimed to characterize the overall characteristics of the paper (Questions numbered 1-3), the research methods used (Q4-5), the type and amount of empirical material (Q6) and additional comments (Q7). For each of the VTs identified we then filled in a VT form detailed below. We also noted a derived measure based on whether there was zero (noted value 'No') or at least one ('Yes') VT forms filled out for each paper. All the sheets and the answers to each question were collected, sorted per paper, in an excel sheet for further analysis.

D. Validity threat form

For each validity issue discussed in one of the papers we used the form summarized in Table III to describe it.

The first two questions on the form (Q1-2) asked the reviewer to describe the validity threat and if the authors gave the threat any 'formal' name according to Table II above. Then two questions (Q3-4) were filled out for each mitigation strategy associated with a validity threat. One questions focused on describing the mitigation strategy itself (in the authors own words) while the other asked the reviewer to estimate which part of the study had been affected by the mitigation. The included parts to be judged were the design, the collected data or the analysis of the results. If the mitigation strategy had not actually been used but was just noted as a possible future strategy it was marked as affecting the 'Future Work'.

The total time needed for reviewing each paper was also noted in the excel sheet that collected all information.

Table II: Summary of 'Paper form' with questions for each analysed paper.

Question	Alternatives
1. Software Engineering Perspective?	Business, Architecture/Technology, Process, Organisation, *Other*
2. Software Engineering Areas?	Requirements Engineering, Design or Architecture, Implementation, Testing or V&V, Management, Metrics or Measurement, People issues or Human factors, Economics, Methodology, *Other*
3. Sub-area(s) within main area?	*List main keywords in order of decreasing importance*
4. Type of methodology or outlook?	Quantitative, Qualitative, Both, Unclear
5. Research Methodology(ies)?	Controlled Experiment, Experiment, Survey, Case study (comparative), Case study (exploratory), Multiple Case Study, Design of Solution, Improvement of Solution, Design of Education, Design of Methodology, Observation study, Grounded Theory Study
6. Data sources and collection methods?	*Describe type as well as number of subjects/groups/companies etc*
7. Comments?	*Free form for any important additional comments*

Table III: Summary of 'Validity threat form' with questions for each validity issue.

Question	Alternatives
1. Summary of validity threat?	*Summarize based on how threat is described in the paper*
2. Classification or naming of threat by author?	*Note according to validity threats listed in Table I*
3. Mitigation strategy for threat?	*Summarize strategy based on how described in the paper*
4. Mitigation strategy affected which part of study?	Design of study, Implementation / Collected data, Analysis of results, Future work

Table IV: Num. of papers for different types of research methodology.

Type of Method	Number	Percentage
Quantitative	30	71.4%
Qualitative	4	9.5%
Both	7	16.7%
Unclear	1	2.4%
TOTAL	**42**	**100%**

IV. Results and Analysis

A. Summary of quantitative data

From the 42 paper forms included in the analysis only one focused on the Business perspective, 11 on the Architecture/Technology perspective, 28 on the Process perspective and four on the Organisational perspective[1]. One paper was focused on Education.

Table IV shows the distribution of papers for the different types of research methodology (Q4 on 'Paper form'). Unfortunately the distribution is very skewed and have too few Qualitative papers for us to be able to compare the two different methodologies.

Table V summarizes the number of papers and their average number of validity threats for each research methodology employed. Overall 34 papers (79.1%) discussed at least one validity threat while 9 papers (20.9%) lacked any such discussion. We can see that interestingly enough there seems to be a difference in the number of validity threats that are discussed in quantitative studies compared to qualitative ditto. If we consider the experiments, controlled experiments and surveys as quantitative methods and consider case studies, multitude case studies, observation study and action research as qualitative (since the latter used grounded theory it seems plausible), the former group considers an average of 5.46 validity threats while the latter considers 3.23. Since we cannot assume the number of validity threats follow a normal distribution and since they are few, we compared these differences in mean with the Wilcoxon rank sum non-parametric statistical test. The difference between them are statistically significant at the $\alpha = 0.05$ level ($p = 0.029$).

Figure 1 shows a boxplot of the number of validity threats for the papers employing qualitative or quantitative research methods, respectively. We can see that the difference is quite clear as confirmed by the statistical test.

The number of mitigation strategies (69) was much lower than the number of validity threats (136) with an average of only 1.60 mitigation strategies per paper and 0.49 per validity threat. So on average only for about half of the validity threats that were discussed did the paper authors also discuss a way to overcome the threat.

Table VI lists the number of mitigation strategies deemed to affect a certain phase of a study. We note that for 5 of the 69 identified strategies it was not possible to judge which phase they affected. For the rest there is no clear trend, other than the fact that many mitigation strategies (26.5%) even if they are discussed are just mentioned as future work and have not affected the results in the reviewed study at all.

[1]Note that a few papers had more than one perspective

Table V: Num. of papers for different research methodologies.

Method	Num. papers	Percentage	Avg. VTs
Experiment	9	21.4%	5.22
Design of Solution	8	19.0%	1.38
Case studies	7	16.7%	4.14
Multiple case studies	4	9.5%	2.25
Design of Improvement	4	9.5%	1.25
Survey	3	7.1%	4.33
Design of Methodology	3	7.1%	1.67
Controlled experiment	1	2.4%	11
Design of education	1	2.4%	3
Observation study	1	2.4%	2
Grounded theory	1	2.4%	2
TOTAL	**42**	**100%**	3.26

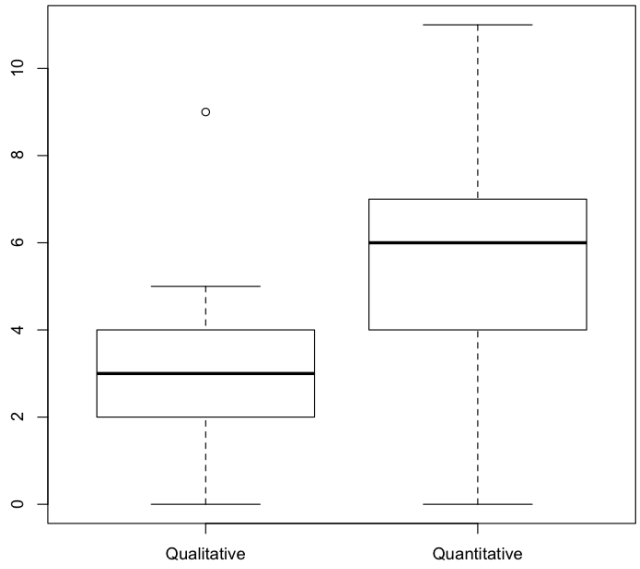

Figure 1: Boxplot of num. of validity threats for different research methods

Table VI: Num. of mititgation strategies that affects each phase.

Affected phase	Num. strategies	Percentage
Design	20	31.2%
Future Work	17	26.5%
Analysis	14	21.9%
Implementation / Data collection	13	20.3%
TOTAL	**64**	**100%**

B. Analysis of validity threats in reviewed papers

Almost no authors had classified their threats based on any of the standard methods or names as presented in Section II. The only widespread name was 'Generalizability' which is the same as External validity.

A fairly common threat was the questioning of the 'quality of the raw data', 'limited sample size' and 'convenience sampling'. Quite common was also the use of different types of biases to describe threats. Example of such biases that were mentioned are the 'researcher intervention bias', 'mono-operation bias', 'mono-method bias', 'researcher bias', and 'vested interest might cause bias'.

Overall it was not clear how to map many of the discussed threats into the formal nomenclature. Most authors used their own naming and did not try to relate to the published literature with lists of threats to be checked. Many authors seemed to consider the validity threats mostly as a post-research walkthrough of limitations with only limited actual effect on the study. However, we need to study the data in more detail, and find a unified framework in which to analyse the many different threats stated before we can investigate this in more detail.

We also investigated how easy it was to find the validity discussion in the reviewed papers. Only 25 of the investigated 43 papers had an explicitly named section discussing validity threats.

V. Discussion

Overall more than 20% of the reviewed papers lacked any analysis of validity threats. This is an alarmingly high figure. Empirical researchers should know that there is always several threats to the validity of any research study and they should be examined both before and while designing the study as well as throughout the other phases of a study.

The average number of validity threats discussed was 3.26 overall, but higher for quantitative studies (5.46) and slightly lower for qualitative studies (3.23) and the lowest for studies that designed a new method, approach or solution. We think this can be explained by the fact that the existing guidelines for validity analysis of software engineering research focus on experiments. It thus becomes more likely that researchers doing quantitative research in the area would eventually get reviewer comments or take preparatory courses etc where they are introduced to these guidelines. The same is not as likely for qualitative researchers which would have to go outside of the software engineering field to find relevant guidelines and support.

It is apparent from our study that the existing terminology for analysing validity threats is not used. Only in a few cases was the terms Internal, Conclusion, Construct or External validity used, even in experiments and other quantitative studies. We propose that one explanation can be the fact that these terms are not more directly linked to the actual elements of the studies, what has been done etc. Rather many researchers seem to describe threats based on their within-study terminology and understanding. The current terminology is more abstract which may limit its uptake and use and thus the level of support it gives.

To remedy this situation we proposed that a new and simpler terminology is introduced that is more directly linked to the actual elements of the study and is adapted to the type of method, data source and analysis method employed. Ideally such a framework should support both quantitative and qualitative methods. We see little reason that this should not be possible as long as variation is allowed along the dimensions discussed above. As a starting point a common and abstract process model of most research studies can be used along the lines presented for case studies in [11] or for experiments in [1]. These two models both have five steps with the same meaning and only slightly varying names.

That the number of mitigation strategies is low compared to the number of threats might seem alarming. However, many threats are stated just as limitations for which there is no way the author could have mitigated the it. A typical example is that the sample size is too small so the results might not be statistically significant. However, the reason often seem to be one of convenience sampling, i.e. the researchers interviewed or performed experiments on the subjects that were available in the studied organisation; they could not get access to any additional subjects.

There also seem to be some room for more extensive validity analysis. For the 25 reviewed papers that had a validity discussion they covered 5.44 threats on average per paper. Since there are typically several threats to each part and phase of a study this number seems low and it seems that there may be a lot of ommission errors. However, it can also be the case that because of length restrictions the researchers have considered more threats but only includes the most severe ones in the actual paper. For these reasons it might be useful to have more complete validity analysis available on home pages for respective papers or similar. Standardized forms for this type of analysis and reporting might be valuable to promote this practice. However, we first need to extend the analysis in this paper to better understand the quality of validity analysis today.

A. Validity threats

There are several validity threats to the design of this study. Our choice of venue is limited to a single venue and a single year. In extending this work we should of course include more venues and more years. This would both give more data and allow better and more detailed statistical analysis as well as allowing a broader coverage of the field of software engineering. Care should be taken to ensure that also quality studies are found and included.

During data collection we mostly used a single researcher to review each paper. Although we tried to mitigate this threat by noting any unclear issues and discuss them together there

is still a higher risk that a single reviewer can be biased and consistently extract the wrong information. For the future we hope to have the resources to have at least two reviewers for each paper. The pilot with co-review of four papers should have helped to create a common understanding though.

Another threat to the data collection is that our chosen categories did not always fit the papers, threats and/or mitigation strategies. Based on the experience from this initial survey we will update the form further to ensure consistent collection and analysis of results.

We see few threats to the numerical analysis, however there are several threats to the analysis of the threat and strategy summaries. Since we had no unified framework with which to classify the threats or strategies our analysis at this stage is only indicative; without statistics on which threats are common and not the analysis is unreliable.

VI. Conclusions

To investigate the current state-of-practice in analysing validity threats among researchers in empirical software engineering we have reviewed all full research papers from the ESEM conference in 2009. More than 20% of these papers contains no discussion of validity threats and the ones that do discuss on average only 5.44 threats. For only half of the discussed threats are any strategy to overcome or mitigate the threat discussed by the paper authors of the reviewed papers. And more than 25% of these mitigation strategies mentioned have not been used in the studies but are just discussed as future work. Furthermore, the situation seems to be worse for qualitative studies than for quantitative, possibly because the little advice there is for validity analysis in software engineering has been presented for experiments. However, even the existing advice seem to have found little use, in particular, no common terminology seemed to be used for validity analysis; researchers use their own terms or use no specific terms at all.

We propose that a common model for the process of conducting empirical research in software engineering is created and that a simpler terminology for validity threats and analysis is adapted to this model. Specific guidelines for different research methods, data sources and collection methods can then be attached to the model and allow adaptation to the specifics of each study. Such a model would also allow for more detailed analysis of which threats are currently analysed and which have been missed. In future work we will developed such a model and framework and apply it to more research papers to validate it as well as deepen our knowledge of the state-of-the-art in validity analysis.

References

[1] C. Wohlin, M. Höst, P. Runeson, M. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in software engineering: an introduction*. Kluwer Academic Publishers, 2000.

[2] J. Maxwell, *Qualitative research design: An interactive approach*. Sage Publications Inc., 2004.

[3] D. Šmite, C. Wohlin, T. Gorschek, and R. Feldt, "Empirical evidence in global software engineering: a systematic review," *Empirical Software Engineering*, vol. 15, no. 1, pp. 91–118, February 2010. [Online]. Available: http://dx.doi.org/10.1007/s10664-009-9123-y

[4] S. Ali, L. Briand, H. Hemmati, and R. Panesar-Walawege, "A Systematic Review of the Application and Empirical Investigation of Search-based Test-Case Generation," Technical Report Simula. SE. 293. Simula Research Laboratory, Norway, Tech. Rep., 2009.

[5] K. Petersen and C. Wohlin, "Context in industrial software engineering research," in *Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement*. IEEE Computer Society, 2009, pp. 401–404.

[6] T. Cook, D. Campbell, and G. Fankhauser, *Quasi-experimentation: Design & analysis issues for field settings*. Houghton Mifflin Boston, 1979.

[7] A. Bryman and E. Bell, *Business research methods*. Oxford University Press, USA, 2007.

[8] N. Mays and C. Pope, "Qualitative research in health care: Assessing quality in qualitative research," *British Medical Journal*, vol. 320, pp. 50–52, 2000.

[9] J. Mason, *Qualitative researching*. Sage Publications Ltd, 2002.

[10] Y. Lincoln and E. Guba, *Naturalistic inquiry*. Sage Publications Inc., 1985.

[11] P. Runeson and M. Höst, "Guidelines for conducting and reporting case study research in software engineering," *Empirical Software Engineering*, vol. 14, no. 2, pp. 131–164, 2009.