Identifying Threats to Validity and Control Actions in the Planning Stages of Controlled Experiments

Amadeu Anderlin Neto and Tayana Uchôa Conte
Instituto de Computação
Universidade Federal do Amazonas (UFAM)
Manaus, AM, Brazil
{neto.amadeu, tayana}@icomp.ufam.edu.br

Abstract—During the planning phase of an experiment, we must identify threats to validity in order to assess their impact over the analyzed data. Furthermore, actions to address these threats must be defined, if possible, so that we can minimize their impact over the experiment’s results. In this paper, we propose the Threats to Validity Assistant (TVA), which is a tool to assist novice researchers in identifying and addressing threats to validity of controlled experiments in Software Engineering. The TVA was evaluated through a controlled experiment which aimed at measuring the effectiveness and efficiency in identifying threats to validity and actions to address them. Sixteen subjects were asked to identify threats to validity and control actions related to an empirical study reported in literature. Our results indicate that the subjects who used TVA were significantly more effective and efficient than the subjects who did not use it.

Keywords – validity, controlled experiment, empirical study

I. INTRODUCTION

Empirical studies have long been applied to provide confidence in assertions about what is true or not true in the Software Engineering (SE) domain [12]. According to Wohlin et al. [14], empirical research is responsible for the maturation of SE.

There are growing numbers of empirical studies to evaluate different SE practices and techniques [9]. New technologies should be compared with existing technologies through experimentation. Thus, we can collect evidence related to the performance of a technology (e.g., benefits, limitations, costs, risks). However, conducting experiments is a complex task [8] because researchers can cause changes in the experiment environment which could influence its outcomes and introduce threats that may compromise its validity.

According to Biffl and Halling [4], every empirical study presents threats to validity (TTVs). These TTVs must be identified and addressed [11]. The identification and mitigation of TTVs are critical activities that require a lot of effort from researchers. TTVs are potential risks that may arise during the planning and execution of empirical studies [14]. Thus, experiments must be carefully planned and executed.

Researchers in Empirical Software Engineering (ESE) have proposed guidelines, checklists, and summaries to support the identification and mitigation of TTVs [7]. However, none of these present relationships between TTVs and control actions (CAs).

Novice researchers have difficulty in identifying and addressing TTVs during the experiment planning. Thus, such problem motivated us to build a conceptual model presenting relationships between TTVs and CAs [1]. The proposed conceptual model shows which CAs can be applied to address a TTV. In addition, it also shows the new TTV that may occur when applying a CA to address a certain TTV.

As a way to facilitate the use of the conceptual model, we developed a tool, the Threats to Validity Assistant (TVA). The main idea is to assist novice researchers in identifying and mitigating TTVs. Therefore, we aim at enabling researchers to: (a) reduce the effect of the TTVs that can compromise the results of empirical studies; and (b) increase the degree of confidence in the obtained conclusions by enhancing the quality of the empirical studies.

The scope of this work focused only in TTVs of controlled experiments. Kampenes [8] defines controlled experiment as “a randomized experiment or a quasi-experiment in which individuals or teams (the experimental units) conduct one or more software engineering tasks for the sake of comparing different populations, processes, methods, techniques, languages, or tools (the treatments)”.

In this paper, we present a controlled experiment which evaluated the TVA. This experiment was conducted in order to measure the effectiveness and efficiency in identifying TTVs and CAs. The quantitative results indicated that the experimental group had better performance than the control group.

This paper is organized as follows. Our approach to assist the identification and mitigation of TTVs is described in Section II, while its evaluation follows in Section III. Then, we present the experiment’s results in Section IV and, in Section V, we explore its TTVs. Finally, Section VI concludes the paper and describes future work.

II. THE APPROACH TO IDENTIFY AND ADDRESS TTVS

We followed a few stages to develop our approach which assists the identification and mitigation of TTVs. In the following subsections, we describe the conceptual model and the TVA tool.

A. The Conceptual Model

The conceptual model shows TTVs and CAs which were described in experiments reported in literature. We selected
these TTVs and CAs through a Systematic Literature Review (SLR). The steps for performing the SLR are described in [1].

We organized the conceptual model as follows. First, we grouped the TTVs with their respective CAs. After that, we grouped only TTVs and CAs which present trade-off. The trade-off is characterized when a CA is applied in order to control a TTV and, by applying such CA, new TTVs may arise.

Figure 1 shows part of the conceptual model. The codes presented in each entity correspond to the ID of the TTV or CAs. These codes follow the structure: ValidityType-TypeOfEntity-Number. For instance, EXT-T01 means the entity is the first (01) TTV (T) to external validity (EXT) identified within the results from the SLR. This TTVs can be controlled by CA EXT-C21 (in which C means control). However, this CA may cause TTV COS-T03 (in which COS means conclusion validity). In order to address this TTV, COS-C04 can be applied. Nevertheless, CA COS-C04 may cause TTV EXT-T01.

![Figure 1. Trade-off between threats to external and conclusion validities.](Image)

As shown in Figure 1, there is a trade-off between validity types [3]. When increasing one type, another type may be decreased [14]. The researchers are responsible for prioritizing the validity types according to their needs. Depending on the experiment, some TTVs are more critical than others [2].

The conceptual model allows a visualization of the ways to mitigate TTVs and consequences that may occur when the researcher chooses a CA, since relationships (trade-offs) between TTVs and CAs are mapped. As way to facilitate the use of the conceptual model, we have developed a tool, which is described below.

![Figure 2. (a) Part of the checklist page in the TVA; (b) Example of the threat list in the TVA.](Image)

**B. The Threats to Validity Assistant**

The Threats to Validity Assistant (TVA) was developed in order to facilitate the use of the conceptual model. We used the Grails\(^1\) framework and the MySQL database to develop the tool. All relationships presented in the conceptual model were mapped to the database within the TVA.

From the list of TTVs obtained from SLR, we derived a checklist. The checklist contains 65 items which aim to capture the experiment context and identify TTVs that may occur. Such items are related to one or more TTVs. Likewise, the TTVs are related to one or more items. Thus, there are relationships between items and TTVs.

Besides the relationships between items and TTVs, there are two relationships between TTVs and CAs. The first one is a relationship in which a CA can be applied to address a TTV. Each TTV can have none, one or more CAs. Otherwise, a CA can control one or more TTVs. The second relationship characterize the trade-offs. A new TTV may arise when a specific CA is applied to control a TTV. Thus, one CA can cause none, one or more TTVs.

Figure 2a shows part of the checklist page where we have selected the items “The data collected can be inaccurate” and “There is the possibility of communication between the subjects during the experiment”. Once the researchers fill in the checklist, the next screen of the TVA shows the list of possible TTVs. For example, Figure 2b shows a list of TTVs which was produced according to the selected items in the previous page (Figure 2a).

Researchers may select one or more CAs to address a TTV, or even none, according to the purpose of the experiment. For instance, Figure 3a shows the CAs to address the TTV “Communication among subjects during the experiment”. When saving the selected CAs, if a CA may cause a new TTV, a warning message appears for the researcher. If the researcher confirms the use of the CA, the new TTV will be included in the list of TTVs on the previous screen (Figure 2b). In our example, we selected “Introduce observers in experimental environment” to address the TTV “Communication among subjects during the experiment”. However, this CA can cause the TTVs “Experimental environment representativeness” and “Subjects act differently when being observed”. Thus, these TTVs will be included in the TTVs list as shown in Figure 3b.

**III. EXPERIMENT DESCRIPTION**

We have designed and executed a controlled experiment to answer the following research question: “Is the identification of TTVs and CAs more efficient and effective when the TVA is used”? To address this question, we have conducted a controlled experiment in order to measure the effectiveness and efficiency in identifying TTVs and CAs, with or without the use of the TVA.

\(^1\)http://www.grails.org/
The definition, the planning (context, variables, hypotheses, selection of subjects, design, and instrumentation) and the operation of the experiment are explained in the following subsections.

A. Experiment Definition

The experiment was defined based on the template proposed in previous work [14] as:

- **Analyze**, Threat to Validity Assistant (TVA) to characterize its effectiveness and efficiency of the researchers
- **For the purpose of**, the identification of TTVs and CAs by students from an Empirical Software Engineering course.
- **With respect to**, its effectiveness and efficiency
- **From the point of view**, from the context of
- **In the context of**, the identification of TTVs and CAs

B. Context Selection

The experiment context was an Empirical Software Engineering (ESE) course offered at UFAM (Federal University of Amazonas) at the north of Brazil. Although participating in the experiment is a mandatory part of the course, the subjects could request that their data were removed from the analysis. The subjects who agreed in participating in the experiment were 10 senior-level (fourth-year) undergraduate Computer Science students, and 6 graduated Computer Science students. The senior-level graduated students enrolled in the course aiming to be prepared at a masters or doctorate degree (more details about the subjects are presented in Subsection E).

The controlled experiment used in this study was selected from the SLR and was based on the following criteria: (1) to describe more than 10 TTVs; (2) to contain TTVs of the four validity types; (3) to present different context used in the pretest activity; and, (4) to follow the experiment process described by Wohlin et al. [14].

Using these criteria, we selected the controlled experiment described by Madeyski [10] (hereafter referred to as Madeyski’s experiment), which contains 23 TTVs and 12 CAs. The context of Madeyski’s experiment was related to testing in software systems.

C. Variables Selection

The independent variable is the treatment applied by the groups in order to identify and address TTVs. Thus, the experimental group used the TVA while the control group used the bibliographical material of the Empirical Software Engineering course, mainly the guidelines described by Wohlin et al. [14]. We chose this book since it was the most cited reference (with about 49 citations within the SLR presented in [1]) in empirical studies to identify and address TTVs.

The dependent variables are the effectiveness and efficiency in identifying TTVs and CAs. Effectiveness and efficiency are based on Conte et al. [5], and are defined as follows.

- **Effectiveness in identifying threats to validity**:\[
\text{Effectiveness} = \frac{\text{number of real threats found}}{\text{total number of threats known}}.
\]

- **Effectiveness in identifying control actions**:\[
\text{Effectiveness} = \frac{\text{number of real actions found}}{\text{total number of actions known}}.
\]

- **Efficiency in identifying threats to validity**:\[
\text{Efficiency} = \frac{\text{number of real threats found}}{\text{time spent}}.
\]

- **Efficiency in identifying control actions**:\[
\text{Efficiency} = \frac{\text{number of real actions found}}{\text{time spent}}.
\]

D. Hypotheses

Using the variables described above, we defined the following null hypotheses:

- \(H_01\) – there is no difference in the effectiveness indicator when identifying TTVs with or without using the TVA;
- \(H_02\) – there is no difference in the effectiveness indicator when identifying CAs with or without using the TVA;
- \(H_03\) – there is no difference in the efficiency indicator when identifying TTVs with or without using the TVA;
- \(H_04\) – there is no difference in the efficiency indicator when identifying CAs with or without using the TVA.
E. Selection of Subjects

The selected subjects were 16 students taking the ESE course. All subjects had lessons about the experiment process described by Wohlin et al. [14]. In addition, all subjects made presentations about validity evaluation in SE experiments and validity types. All subjects signed a consent form and filled out a characterization form that measured their expertise with controlled experiments.

We also applied a pretest activity to measure more directly the subjects’ expertise. The pretest activity consisted in the identification of TTVs and CAs from a paper reporting a controlled experiment. In this activity, the subjects did not use any support. We selected the controlled experiment described by Thelin et al. [13]. This paper was retrieved by the SLR [1], which described a controlled experiment in the context of software inspection.

The pretest activity was conducted simultaneously with all subjects in a classroom. The pretest activity consisted in identifying and addressing TTVs without any support. The subjects had 90 minutes to perform the task. At the end, a researcher had analyzed the answers and assigned a grade to the results of the pretest activity.

Based on the results of the characterization form and the pretest activity, the subjects were divided into groups of: low (L), medium-low (ML), and medium (M) knowledge. Low knowledge means that a person identified one to four real TTVs and CAs, and does not have experience. Medium-low knowledge means that a person identified five to eight real TTVs and CAs, and has little experience (i.e., participated in an experiment as subject). Finally, medium knowledge means that a person identified more than eight real TTVs and CAs, and has little experience. In order to reduce the bias of having more experienced subjects in one or another treatment (with or without using the TVA), we equally and randomly distributed the subjects, balancing both groups.

F. Design of the experiment

The design is one factor, with two treatments. The assignment of the treatments was randomized. Each group consisted of 8 subjects and each used only one treatment, either the TVA or the guidelines proposed by Wohlin et al. [14].

G. Instrumentation

The instruments used in this experiment were: characterization and consent forms, guidelines described by Wohlin et al. [14], data collection form (control group, since the TVA automates such process), TVA, instructions to use TVA (presented on the initial screen of the tool), follow-up questionnaire (experimental group, since it was used to gather improvement opportunities), excerpts from Madeyski’s experiment in the original language (English) and a translated version (Portuguese). These excerpts were: the experiment context, treatments details, metrics details, and experiment planning. The translation was carried out by a researcher and reviewed by another researcher with an advanced English level. All instruments were validated by the authors.

H. Experiment Operation

The subjects were divided into two classrooms, one for the experimental group and another for the control group. All subjects had up to 90 minutes to carry out the experiment. During the experiment, the control group had to fill out the data collection form as they found TTVs or CAs. In addition, the control group used the checklist presented in [14]. On the other hand, the experimental group used the TVA. When the subjects from the experimental group finished the task, they had to fill out the follow-up questionnaire.

After finishing the experiment, we created a list containing all different TTVs without duplicated ones. Likewise, we created a unique CAs list. Both lists were based on TTVs and CAs described in the validity section of Madeyski’s experiment. A total of 23 TTVs and 12 CAs were found in the TTVs section of the paper. However, there were TTVs that were not described in the validity section. Thus, an independent experienced researcher evaluated Madeyski’s experiment and found 23 new TTVs and 29 new CAs. Therefore, a total of 46 TTVs and 41 CAs were identified. Other TTVs and CAs that were found by the subjects and that were not already included in the lists, were classified as false positives (in order words, they were not present in Madeyski’s experiment). In the next section, we present the results of the controlled experiment.

IV. Results

We obtained the quantitative data from the data collection form resulting from the experimental task. After creating the list of TTVs and the list of CAs, we counted the number of TTVs and CAs. We also considered the time used by each subject to carry out the experimental task. Table I presents both the results per subject and per treatment.

We have performed a statistical analysis using the statistical tool SPSS v 20.0.0 and α = 0.05. This choice of statistical significance was motivated by the small sample size used in this experiment [6]. In order to compare the effectiveness and efficiency of both samples, we used the Mann-Whitney non-parametrical statistic method. This test is equivalent of the t-Student parametrical test (for further information, please see [14]). It was used because we had two groups to compare, different subjects in each condition, and no assumption about the data distribution. We also used the boxplots graph to facilitate the visualization of results.

The boxplot graph which compares the effectiveness indicator for the identification of TTVs and CAs is shown in Figure 4. When analyzing the graph, we can see that the mean from the experimental group is higher than the mean from the control group for both indicators. Thus, the group that used the TVA was more effective than the group that did not use it. Also, the comparison using the Mann-Whitney statistic method showed that there was significant difference between the groups (p = 0.001 – for TTVs; and, p = 0.005 for CAs). These results suggest that the use of the TVA provided different effectiveness related to the identification of TTVs and CAs in controlled experiments; and that this difference is significant from a statistical point of view. These results reject the null hypothesis H₀₁ and H₀₂.
### TABLE I. QUANTITATIVE RESULTS PER SUBJECT AND TREATMENT.

<table>
<thead>
<tr>
<th></th>
<th>Control group (guidelines)</th>
<th>Experimental group (TVA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expertise Level</td>
<td>M  ML ML ML ML ML ML ML ML</td>
<td>M ML ML ML ML ML ML ML</td>
</tr>
<tr>
<td>Time (min)</td>
<td>90 88 75 45 87 87 80 82</td>
<td>70 60 47 62 53 63 52 40</td>
</tr>
<tr>
<td>Total TTVs Found</td>
<td>6 4 7 12 14 8 4 6</td>
<td>18 42 33 50 28 15 35 14</td>
</tr>
<tr>
<td>Real TTVs Found</td>
<td>3 3 7 6 8 8 3 4</td>
<td>13 35 22 37 20 12 25 13</td>
</tr>
<tr>
<td>False Positive TTVs</td>
<td>3 1 0 6 6 0 1 2</td>
<td>5 7 11 13 8 3 10 1</td>
</tr>
<tr>
<td>Effectiveness in Identifying TTVs %</td>
<td>6.52 6.52 15.22 13.04 17.39 17.39 6.52 8.70</td>
<td>28.26 76.09 47.83 80.43 43.48 26.09 54.35 28.26</td>
</tr>
<tr>
<td>Efficiency in Identifying TTVs</td>
<td>2.00 2.05 5.60 8.00 5.52 5.52 2.25 2.93</td>
<td>11.14 35.00 28.09 35.81 22.64 11.43 28.85 19.50</td>
</tr>
<tr>
<td>Total CAs Found</td>
<td>6 4 7 12 10 8 4 5</td>
<td>10 32 17 49 18 6 20 8</td>
</tr>
<tr>
<td>Real CAs Found</td>
<td>1 3 5 3 6 8 1 2</td>
<td>5 26 12 29 12 5 18 6</td>
</tr>
<tr>
<td>False Positive CAs</td>
<td>5 1 2 9 4 0 3 3</td>
<td>5 6 5 20 6 1 2 2</td>
</tr>
<tr>
<td>Effectiveness in Identifying CAs %</td>
<td>2.44 7.32 12.20 7.32 14.63 19.51 2.44 4.88</td>
<td>12.20 63.41 29.27 70.73 29.27 12.20 43.90 14.63</td>
</tr>
<tr>
<td>Efficiency in Identifying CAs</td>
<td>0.67 2.05 4.00 4.00 4.14 5.52 0.75 1.46</td>
<td>4.29 26.00 15.32 28.06 13.58 4.76 20.77 9.00</td>
</tr>
</tbody>
</table>

The boxplot graph which compares the efficiency indicator for the identification of TTVs and CAs is shown in Figure 5. According to the graph, the experimental group was more efficient than the control group, finding more TTVs and CAs in lesser time. Also, the Mann-Whitney statistical test shows that the difference between the groups is statistically significant (p = 0.001 – for both indicators). Thus, the use of the TVA provides better efficiency regarding the identification of TTVs and CAs. These results reject the null hypothesis H₅ and H₆.

### V. THREATS TO VALIDITY

The following subsections present the TTVs considered in this controlled experiment. These TTVs have been grouped in four main categories: internal, external, conclusion and construct.

#### A. Internal Validity
- **Individual differences among the subjects.** In order to address this TTV, we divided the subjects in balanced groups, according to the results from the characterization form and the pretest activity.

- **Language of the study object was different from the native language of the subjects.** In order to address this TTV, we translated the study object to the native language of the subjects. The translation was done by a researcher and reviewed by another researcher with high proficiency in English. Both documents (original and translated) were given to the subjects.

#### B. External Validity
- **Representativeness of subjects.** The students are representative of the target population (novice researchers), but the results cannot be generalized. This experiment should be replicated for other samples and contexts.

- **Representativeness of artifact.** In order to avoid this TTV, we selected an experiment reported in literature. However, new experiments should be conducted with study objects from other contexts.
C. Conclusion Validity

- **The small number of data points.** However, small sample sizes are a known problem in Software Engineering which is difficult to overcome [5]. Thus, the data extracted from this controlled experiment can only be considered indicators and not conclusive.

- **Heterogeneous characteristic of the subjects.** This TTV was not considered because all students do not have experience in the planning of controlled experiments. Thus, the sample was homogeneous.

D. Construct Validity

- **Biased judgment of real TTVs and CAs.** In order to mitigate this TTV, an independent researcher evaluated Madeyski’s experiment and judged which were real TTVs or CAs. However, only one independent researcher might not be enough to classify TTVs and CAs.

- **Only one study object was used.** Therefore, the results may depend on the study object used. In order to mitigate this TTV, we plan to conduct new controlled experiments with different study objects, to enhance the validity of our results.

- **Subjects may behave differently, because they were observed during the experiment execution.** In order to avoid this TTV, subjects were observed during the pretest activity. Thus, it is expected that the subjects adapted to the environment. However, the observation impacts cannot be measured.

In summary, we conclude that the existing TTVs were not considered large in this study. Also, we carried out CAs in order to reduce their impact.

VI. CONCLUSION AND FUTURE WORK

This paper has presented a controlled experiment to answer the following research question: “Is the identification of TTVs and CAs more efficient and effective when the TVA is used?” In order to answer this question we have measured the results of the TVA in terms of effectiveness and efficiency.

Results showed that the use of the TVA provided better effectiveness and efficiency related to the identification of TTVs and CAs. The quantitative analysis using the Mann-Whitney statistic method showed that the effectiveness and efficiency indicators are better, when using the TVA, and that there is significant statistical difference. Due to the small sample used and the number of carried out experiments (only one), these results are not conclusive but indicate that the use of the TVA enhance the effectiveness and efficiency of the identification of TTVs and their CAs.

We intend to perform new empirical studies to validate the results, including subjects with high experience in planning controlled experiments. Thus, we will obtain evidence if the TVA is useful for experienced researchers, while evaluating if experienced researchers agree with our categorization and relationships of TTVs and CAs.

We hope that our findings will be useful in the promotion and improvement of the current practice of experiments planning. We also hope that the proposed tool aids researchers in the validity evaluation of their empirical studies, improving the confidence in their results and reducing the effort during planning stages.

ACKNOWLEDGMENT

We would like to acknowledge the financial support granted by FAPEAM process PAPE 032/2013.

REFERENCES


