Reporting Guidelines for Simulation-Based Studies in Software Engineering

Breno Bernard Nicolau de França, Guilherme Horta Travassos
Federal University of Rio de Janeiro
PESC / COPPE
Rio de Janeiro, Brazil
bfranca@cos.ufrj.br, ght@cos.ufrj.br

Abstract—**BACKGROUND: Some scientific fields, such as automobile, drugs discovery or engineer have used simulation-based studies (SBS) to faster the observation of phenomena and evolve knowledge. All of them organize their working structure to perform computerized experiments based on explicit research protocols and evidence. The benefits have been many and great advancements are continuously obtained for the society. However, could the same approach be observed in Software Engineering (SE)? Are there research protocols and evidence based models available in SE for supporting SBS? Are the studies reports good enough to support their understanding and replication? AIM: To characterize SBS in SE and organize a set of reporting guidelines aiming at improving SBS' understandability, replicability, generalization and validity. METHOD: To undertake a secondary study to characterize SBS. Besides, to assess the quality of reports to understand the usually reported information regarding SBS. RESULTS: From 108 selected papers, it has been observed several relevant initiatives regarding SBS in software engineering. However, most of the reports lack information concerned with the research protocol, simulation model building and evaluation, used data, among others. SBS results are usually specific, making their generalization and comparison hard. No reporting standard has been observed. CONCLUSIONS: Advancements can be observed in SBS in Software Engineering. However, the lack of reporting consistency can reduce understandability, replicability, generalization and compromise their validity. Therefore, an initial set of guidelines is proposed aiming at improving SBS report quality. Further evaluation must be accomplished to assess the guidelines feasibility when used to report SBS in Software Engineering.**

Keywords-Computer Simulation, Simulation Studies, Software Engineering, Systematic Review, Guideline.

I. INTRODUCTION

Simulation-based studies (SBS) are claimed to reduce the risks, time and costs of experiments, considering they usually run in virtual environments. Besides, it is also claimed they facilitate the replication of experimental studies and allow the testing of hypotheses before their implementation into *in vivo* or *in vitro* studies [1], anticipating the effects of such implementation. These are believes that could influence the interest on SBS in the Software Engineering (SE) community.

Simulation is one of the topics researched through Systematic Literature Reviews (SLRs). For instance, Zhang et al [2] performed a SLR aiming at tracing the evolution in Software Process Simulation research from 1998 to 2008. Among their contributions, the authors emphasize: "*Categories for classifying software process simulation models; Research improving the efficiency of Software Process Simulation Models is gaining importance; Hybrid process simulation models have attracted interest as a possibility to more realistically capture complex real-world software processes*".

Because of our broader interest in SBS, we have undertaken a secondary study with the goal of characterizing the reported SBS in the SE technical literature [5]. Looking for the anticipation of results before implementing software solutions, we started to look for a simulation-based alternative regarding software architecture tradeoff analysis considering distinct design options. Besides, we looked for a methodology to support the development of valid simulation models, and to improve guidance in the planning and execution of SBS. As our secondary study has a characterization purpose, although strictly following the systematic review approach, there is no previous baseline allowing comparisons. Thus, we call this review a *quasi*-Systematic Review [3].

The main research question is *"How have the different simulation approaches presented in the technical literature been applied in simulation-based studies in Software Engineering?"*. The search string structure follows PICO strategy [4]. The 9 control papers were captured in previous *ad-hoc* literature review. Scopus, Ei Compendex and Web of Science (ISI Knowledge) were selected as source of information. From 1492 entries (906 from Scopus, 85 from Web of Science and 501 from Ei Compendex) with 546 duplicates, only 108 papers were selected after applying all exclusion criteria and detailed reading to support information extraction for analysis. See details in [5].

Our results complement Zhang et al [2] ones. Several SBS features were identified considering the context of SE, including 19 simulation approaches, 17 software engineering domains, 28 simulation models characteristics, 9 procedures for assess the quality of simulation models and 22 output analysis instruments. Besides, the most dominant simulation approaches in SE are system dynamics (appearing in 37 models and 51 papers) and discrete-event (appearing in 17 models and 18 papers). Hybrid models also mostly present combinations of these two approaches. The reasons for distinct findings among

the studies may be related to the different research questions, considering they share similar population of primary studies.

In several papers, the simulations models are not detailed. Researchers only mention the underlying simulation approach. So, 28 characteristics explicitly mentioned within model descriptions have been identified. The results comply with the technical literature [6] regarding software engineering domains. Software process (18 models in 19 papers) and software project (24 models in 35 papers) are the most dominant SE domains in SBS. Software Architecture and Design domain aggregates simulation models concerned with design issues for different classes of systems, such as: fault-tolerant systems, embedded systems and real-time systems, under the quality attributes perspective such as reliability and performance. This domain is characterized by simulating mostly the product (design specification) than the design process. The most dominant V&V procedure is the "Comparison against actual results", used in 10 simulation models found in 17 papers. It represents an interesting way to verify the model "accuracy" in terms of its output results, but many other threats to study validity should be evaluated leading the combined use of other procedures.

Despite all relevant information gathered with the *quasi* SLR, some issues were also observed in the reported SBS in Software Engineering. For instance, most of the reports do not present comprehensive information regarding the research protocol, how the simulation model has been built and evaluated, which data were used, among others. The reported results are usually specific, making hard their generalization and comparison. Actually, it is not possible to observe any reporting consistency among the evaluated studies' results. The lack of such information impacts the quality of SBS. Therefore, this paper offers an initial set of guidelines containing a list of needed information when reporting simulation based studies in SE. It aims at contributing to improve understandability, replicability, generalization and validity of such studies.

The rest of this paper is organized as follows: section 2 presents a simulation studies profile; section 3 presents quality assessment performed over the SBS captured in the review; section 4 presents the guidelines offering explanations for each of the proposed items; and section 5 presents the final remarks.

II. SIMULATION STUDIES PROFILE

From the 108 selected papers, only 57 present descriptions regarding the primary studies although it seems to exist a misunderstanding on classifying the used study strategies. Most of the reported studies are examples of use (assertions or informal studies [7]), with no systematic methodology reported to support their planning, execution or analysis. We adapted existing study strategies concepts from the glossary of terms (lens-ese.cos.ufrj.br/wikiese) and ontology (lens.cos.ufrj.br/esee) offered by eSEE – experimental Software Engineering Environment [8] to classify the observed simulation studies as: survey (retrospective data), case study (current real data) or controlled experiments. These surveys are not designed as a collection of expert's opinion using forms as instruments, but surveying historical project data through simulation trials using simulation models as instruments. Actually, these surveys are closer to what the technical literature call "simulation experiments" [9] differing from "controlled experiments".

Survey studies represent 57.9% of the studies. The most dominant procedure (64.9%) is "Variation of input parameters to observe the impact on output variables". It was not possible to observe any trend regarding input parameter, although ad-*hoc* constant approaches seems to be more frequent (45.6%). Model calibration is usually not applied or reported (54.3%).

The majority of studies use system dynamics models (36 studies). Furthermore, just a few SBS replications, most of them by the same authors, were identified. A possible reason for lack of replication can be the insufficient information about the simulation models. Similarly, the Software Project Management domain is the most studied domain using SBS (25 studies). Besides, the combination of System Dynamics models with Software Project Management domain can be observed in most of the simulation studies in this review.

The purpose and goals of simulation models and studies are not clearly defined in the analyzed papers. It is very hard to find structured research questions, hypotheses or even a *GQM (Goal-Question-Metric)* template, for example, describing the main investigation goal. In all cases it is difficult to identify the experimental design, including the descriptions regarding control and treatment groups. In few cases, it is possible to identify factors and response variables for the studies where the output data is presented usually by charts. It is not easy to find answers to questions such as *what are the treatments for each experimental factor? What are the model input parameters? Do they remain constant? What were their initial values for each simulation trial?* Information as the number of simulation trials is seldom reported. When reported, criteria explaining why such number is used are missing. By not addressing these issues, the replication and auditing of these studies represent unfeasible tasks, and it is hard to compare the results or benchmarking simulation models, since there is no baseline. The works in [11] and [12] are exceptions into this context.

III. QUALITY ASSESMENT

Considering the aforementioned issues, we assessed the reports quality to better understand what kind of information could be missing and look for a possible cause for it. The assessment criteria (Table I) applied to all 108 research papers have direct linkage with the information extraction form fields [5].

TABLE I. QUALITY EVALUATION CRITERIA

Criteria	Value
Does it identify the underlying simulation approach?	*1 pt*
Does it explicitly mention the simulation model purpose?	*1 pt*
Does it explicitly mention the study purpose?	*1 pt*
Does it identify the SE domain the study was undertaken?	*1 pt*
Does it explicitly mention any tool support?	*0.5 pt*
Does it mention the simulation model characteristics?	*0.5 pt*
Does it present a classification or taxonomy for the characteristics?	*0.5 pt*
Does it present the simulation model advantages?	*0.5 pt*
Does it present the simulation model disadvantages?	*0.5 pt*
Does it present any V&V procedure regarding the simulation model?	*1 pt*
What are the statistical instruments used in the output analysis?	*1 pt*
Does the study strategy used in the SBS is identified?	*0.5 pt*
Does it identify the experimental design of the simulation study?	*1pt*

The mean score is 6.16 (in a [0-10] scale) and standard deviation is 1.41. Such scores can be interpreted as lack of important information in the reports. Also, the low standard deviation shows that even the more complete reports do not offer much more information. The percentage of papers reporting each type of information is shown in Fig. 1. Darkest bars indicate those ones strictly connected with SBS issues.

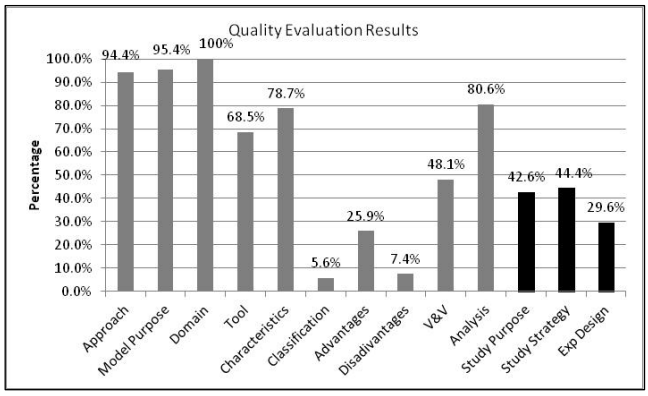

Figure 1. Quality Evaluation Results

All 108 papers present enough information to identify the SE domain in which the simulation study is taking place. However, different from expected, just six occurrences (5.6%) of study classification were found. It does not mean a true classification, but groups of characteristics by which the authors believe that can be possible to organize the simulation models. However, these papers do not discuss this issue, they just mention it. Less than half of the papers mention, introduce or discuss V&V procedures concerned with simulation models. Information about the studies is also very scarce. Just 48 research papers present a clear goal (what is the prime information to be reported) and only 32 research papers present some information regarding the experimental design, which often is not fully described. Dependent and independent variables can be found in some of them without explanation about their arrangement or number of simulation runs.

IV. Reporting Guidelines

Software Engineering community has already published some guidelines for experimental studies. Kitchenham et al [13] propose preliminary guidelines to assist researchers, reviewers, and meta-analysts in designing, conducting and evaluating empirical studies in general. In that work, authors mentioned the need for specific guidelines for different types of studies. Later, addressing this issue, Höst and Runeson [14] proposed specific guidelines for case studies.

Simulation-based studies have also issues to be addressed. In this way, we looked for guidelines for simulation studies in SE, but we could not find one. So, statistics and medicine research areas were visited to look for such guidelines. Ören presents a set of concepts and criteria to assess the acceptability of simulation studies in general [10]. Balci presents guidelines for successful simulation studies [15]. In [16], several experimental designs issues are discussed. In medicine, Burton et al [17] present and discuss a checklist highlighting important issues for designing simulation studies. Essentially, studies should not be different from their reports [14]. Thus, every planned and executed procedure including the decisions taken during the experimentation process must be explicit.

A. Goals and Scope

The goals definition is the first step. It needs to be described in a clear way, leaving no doubt about what questions to answer, in the same way it occurs with others study strategies. For instance, it is likely to find goals definition using *GQM* approach and it seems to be enough for those strategies such as regular surveys, case studies or controlled experiments. Besides, the scope shall be explicitly stated, establishing boundaries for the research area, problem, domain, and type of systems or processes under investigation.

The SBS goals must match the capabilities of the simulation model. In other words, the simulation model should be able to support the answers for the research questions through the output data, and its input parameters should allow the desired scenario configuration. The study must also investigate one or more hypotheses. It should be reported, clarifying the null and alternative hypotheses. Also, it is useful to discuss how such hypotheses were raised, describing the rationale or theory from where they came from.

B. Model Description and Validation

When describing the simulation model, it is relevant to detail its variables and their relationships, as well as its input parameters and the range of values for each one. It is useful, for instance, to complement the information regarding the experimental design. Such description must include the underlying simulation approach. It is important to clarify such approach from the characterization point of view. The abstraction and execution mechanisms are immediately understood by presenting the simulation approach. For instance, on describing a system dynamics model, it is possible to infer how simulations are executed; the stocks and flows modeling abstractions; the causal relationships and feedback loops are also expected to be shown.

The SBS validity is highly impacted by the simulation model validity. If the used model is not valid, invalid results will be obtained regardless the other possible threats to the study validity. Previous reports or research papers presenting evidences regarding the simulation model validity must be described or referenced. In case of absence of such validation references, verification and validation procedures should be performed to assure the model validity, reporting the results as well as the decisions that guided the validation process.

From many simulation models found in Software Engineering, just a few report performance measures. Bias, accuracy and confidence are several times not reported. The importance of such measures relies in the possibility of using them as benchmark criteria in order to compare and choose more accurate simulation models. Such information also brings credibility to the simulation study. Burton et al discuss on how to calculate such measures [17].

C. Simulation Scenarios and Subjects

When investigating a system or a process through SBS it is common to make use of scenarios [18]. The relevance and adequacy of each chosen scenario is important to be described.

To choose and report most representative scenarios, including those ones to both check best and worst cases, can help to foresee the behavior in usual circumstances and in exceptions.

Simulation-based studies may be performed as *in virtuo* or *in silico* categories [1]. In general, SBS makes use of virtual environments. However, it is possible to use individuals or computer programs as subjects. When reporting SBS the study category must be made explicit. Besides, subjects' characterization ought to take place since it can influence the interpretation of *in virtuo* results. The description of subjects' assignment process to the experimental unities must be considered. With computerized subjects, the description of their behavior, configuration parameters and process of assignment must be informed.

D. Experimental Design

Balci mentions four techniques for the design of experiments [15]: *Response-surface* for maximization or minimization of the response variables values by optimizing the combination of parameters values; *Factorial Design* for the determination of the effect of input variables on response variables; *Variance reduction techniques* to obtain better statistical accuracy for the same number of simulations; *Ranking and selection techniques* for comparison of alternative systems (or system configurations). From these four techniques, only factorial design was found in our review regarding SBS in Software Engineering. It can not be claimed all SBS in SE apply this design, but these were the only ones reporting their experimental design with enough detail.

Basically, experimental design issues involve the arrangement of independent variables or factors and the respective treatments definition for each factor. Here, it is clear the importance of the model description and its variables and input parameters. Once they are described, the experimental design can be easily understood and many issues are solved. As observed in [15], system variants may be exploited by different values and types of system parameters, input variables and behavioral relationships, since they constitute the statistical design factors.

Control and treatment groups must be identified when performing controlled experiments using simulation models as instruments. For instance, validated models under known conditions can be assumed as control and the new model (or new versions) to be evaluated or experimented (under the same conditions) can be assumed as the treatment. Another possibility is to use distinct datasets as factors and to fix the simulation model. In this way, different calibrations representing the different simulation scenarios can be compared and should be reported.

E. Number of Runs and Storage of Experimental Trials

The number of simulation runs must be based in the selected simulation scenarios or in the experimental design. Each selected scenario consists in an arrangement of experimental conditions where possible factors are assigned to one specific treatment. The more simulation scenarios are involved in the study, the more simulation runs are needed. For instance, factorial designs usually require one simulation run for each combination of factors and treatments. For a discussion on how to determine the number of simulation runs, based on factorial designs, we encourage to read [11] [16].

For stochastic simulation, the use of random variables should be taken into account since a confidence interval must be estimated from the sample size to determine the number of simulation runs. Such calculation can be found in [17].

SBS involving multiple trials and runs often need to use the information of each trial for output analysis. Means, standard deviations, and other measures are likely to be applied to summarize the whole simulation run (including all trials) and to determine confidence intervals. The report should contain how these measures are stored and used in the analysis.

F. Data Support

When planning SBS it is important to check the availability of supporting data and determine their type: real-system or simulated data. If simulated data has been adopted, some evidence must be presented in order to guarantee the validation of such data, i.e., the report should answer questions like "*How far the simulated data is from real-system data?*" and show indicators of this gap. Here, statistical tests can be applied to verify how close both samples could be.

Simulation models often require time-sensitive data. Hence, in order to avoid biased observations and to be risk exposed (i.e. undetected seasonal data); the data collection time period must represent both transient and steady-state behaviors.

Data collection should be planned to avoid also measurement mistakes, promoting the collection of data as soon as they are make available. After collection, quality assurance procedures ought to take place in order to verify their quality. If the simulation model can be calibrated, it is important to report if it was or not calibrated, including the procedure used to accomplish the task and its results.

G. Tool Support or Simulation Package

The tool support used to automate the SBS is other important feature to be reported. The simulation package must support not only the underlying simulation approach, but also experimental design and data analysis. No research paper was identified in our dataset containing details on how does the simulation package is used and how this instrument works in the study execution.

Simulation packages often differ in how they implement the simulation engine mechanism. So it is possible to get different results depending on how the engine is implemented. Moreover, the process used to translate the conceptual simulation model description to the simulation language offered by the package should be reported. Information concerned with how this translation was performed and if any model characteristic could not be implemented due technological constraints must be presented. In stochastic models, the report of random number generators and how the starting seeds were selected are fundamental.

H. Threats to Validity

SBS reports should, as any other empirical study, discuss the threats to the study validity. However, we will not discuss here the same conventional studies threats of validity, though

they can be applied to simulation studies too. Rather, we concentrate our perspective on threats to validity regarding implications of experimenting with simulation models. All of the common aspects of experimental validity are strongly related to the simulation model validity. It should be valid to assure the study does represent actual system scenarios.

Garousi et al [19] consider that model validity is mainly affected by three factors: (a) proper implementation of cause-effect structures, (b) proper representation of real-world attributes by model parameters, and (c) proper calibration. So, each of these three factors is associated to an aspect of validity: (1) structural validity: related to the building blocks and elements of the corresponding real-world process capture; (2) parametric validity: verification of model parameters suitability for the instrumentation and analysis of the real-world process, and; (3) empirical validity: related to model calibration using data from the corresponding real-world process.

Raffo [20] mentions the model validity in three other aspects: face, output and scenario validity. Face validity tests the fit between the model structure and the real system essential characteristics, often performed by system experts. Output validity verifies the output data accuracy; it involves performance measurements and/or statistical tests. Finally, scenario validity evaluates the meaningfulness of simulation scenarios used in the study, also performed by experts.

External and conclusion validity must be accomplished by the application of adequate statistical tests over the model outputs. However, the conclusion validity is also related to sample size, number of simulation runs, model coverage and representativeness of simulated scenarios for all possible situations.

V. FINAL REMARKS

Simulation studies have been performed in SE since the 80's. Although, the SE field has been evolved in number of proposed simulation models and studies, there are still some issues needing to be addressed. For instance, the study planning, model validity assurance before performing studies and procedures to perform proper output analysis have been taking place with lack of rigor. Except for a few studies presenting a systematic way of doing one or another activity, most of SBS in software engineering are supported *ad-hoc*, as it has been indicated by the results of this *quasi* SLR. Then, this paper presented an initial set of guidelines to promote more information visibility regarding SBS. However, it represents an initial step towards a more comprehensive organization to support the report of SBS. For instance, some dimensions regarding the cost, effort and computer environments involved in the SBS are not completely considered yet.

Hence, we believe methodologies for SBS including model validity assurance and output analysis represent research challenges for those interested on simulation applied in SE. As well, more discussions regarding concrete evidence-based methods for developing simulation models in SE are necessary. It represents an important step to strength SBS in the field. The future in SE also depends on our capacity of accelerating observations and knowledge acquisition. Simulation can play an important role towards it.

ACKNOWLEDGMENT

This research is part of eSEE project and authors thank CNPq for the financial support. We recognize and thank Dr. Marco Antonio Araújo as the second researcher supporting the selection phase of this research protocol trial.

REFERENCES

[1] G. H. Travassos, M. O. Barros, "Contributions of in virtuo and in silico experiments for the future of empirical studies in software engineering," WSESE03. Fraunhofer IRB Verlag, Rome, 2003.

[2] H. Zhang, B. Kitchenham, D. Pfahl, "Reflections on 10 years of software process simulation modeling: A systematic review," LNCS, vol. 5007, pp. 345-356, 2008.

[3] G. H. Travassos, P. M. Santos, P. G. Mian, A. C. Dias Neto, J. Biolchini, "An environment to support large scale experimentation in software engineering," Proc. 13th IEEE ICECCS, pp. 193-202, 2008.

[4] M. Pai, M. McCulloch, J. D. Gorman, "Systematic reviews and meta-analyses: An illustrated,step-by-step guide," The National Medical Journal of India, vol. 17, n. 2, 2004.

[5] B. B. N. França, G. H. Travassos, "Are We Prepared for Simulation Based Studies in Software Engineering yet?," Proc. IX Experimental Software Engineering Latin American Workshop, Buenos Aires, Argentina, 2012.

[6] B. Stopford, S. Counsell, "A framework for the simulation of structural software evolution" ACM Transactions on Modeling and Computer Simulation, vol. 18, 2008.

[7] M. V. Zelkowitz, "Techniques for empirical validation," Empirical Software Engineering Issues, LNCS, vol. 4336, Springer-Verlag Berlin Heidelberg, pp. 4 – 9, 2007.

[8] Lopes, V. P. ; Travassos, G. H. . Knowledge Repository Structure of an Experimental Software Engineering Environment. In: Proc. of XXIII SBES. IEEE Computer Society, v. 1. p. 32-42, 2009.

[9] A. Maria, "Introduction to modeling and simulation," Proc. WSC, 1997.

[10] T. I. Ören, "Concepts and criteria to assess acceptability of simulation studies: a frame of reference," Simulation Modeling and Statistical Computing, vol. 24, n. 4, pp. 180-189, April 1981.

[11] D. X. Houston, S. Ferreira, J. S. Collofello, D. C. Montgomery, G. T. Mackulak, D. L. Shunk, "Behavioral characterization: Finding and using the influential factors in software process simulation models," Journal of Systems and Software, vol. 59, pp. 259-270, 2001.

[12] W. Wakeland, R. H. Martin, D. Raffo, "Using design of experiments, sensitivity analysis, and hybrid simulation to evaluate changes to a software development process: A case study," Software Process Improvement and Practice, vol. 9, pp. 107–119, 2004.

[13] B. Kitchenham, S. L. Pfleeger, D. C. Hoaglin, K. El Emam, J. Rosenberg, "Preliminary Guidelines for Empirical Research in Software Engineering," IEEE Trans. Soft. Eng., vol. 28, pp. 721-734, Aug 2002.

[14] M. Höst, P. Runeson, "Guidelines for conducting and reporting case study research in software engineering," Empir Software Eng, vol. 14, pp. 131–164, 2009.

[15] O. Balci, "Guidelines for successful simulation studies," Proc. Winter Simulation Conference, pp. 25-32, 1990.

[16] J. P. C. Kleijnen, "Statistical design and analysis of simulation experiments," Informatie, 17, no. 10, pp. 531-535, Oct. 1975.

[17] A. Burton, D. G. Altman, P. Royston, R. L. Holder, "The design of simulation studies in medical statistics," Statistics in Medicine, vol. 25, pp. 4279-4292, 2006.

[18] M. O. Barros, C. M. L. Werner, G. H. Travassos, "Applying system dynamics to scenario based software risk management," International System Dynamics Conference, Bergen, Norway, 2000.

[19] V. Garousi, K. Khosrovian, D. Pfahl, "A customizable pattern-based software process simulation model: design, calibration and application," SPIP, vol. 14, pp. 165 – 180, 2009.

[20] D. Raffo, "Software project management using PROMPT: A hybrid metrics, modeling and utility framework," Information and Software Technology, vol. 47, pp. 1009-1017, 2005.